Shooting our hard drive into space and other ways to practise responsible development data

ways to practise responsible development data

CONTENTS

introduction

about this book

am i working with data?

the world of data: opportunities and risks

why responsible data

misconceptions and common myths

how to use this book

about this book

This book is offered as a first attempt to understand what responsible data means in the context of international development programming. We have taken a broad view of development, opting not to be prescriptive about who the perfect "target audience" for this effort is within the space. We also anticipate that some of the methods and lessons here may have resonance for related fields and practitioners.

We suggest a number of questions and issues to consider, but specific responsible data challenges will always be identified through individual project contexts. As such, this book is not authoritative, but is intended to support thoughtful and responsible thinking as the development community grapples with relatively new social and ethical challenges stemming from data use.

This book builds on a number of resources and strategies developed in academia, human rights and advocacy, but aims to focus on international development practitioners. As such, we touch upon issues specifically relevant to development practitioners and intermediaries working to improve the lives and livelihoods of people.

The group of contributors working on this book brings together decades of experience in the sector of international development; our first hand experiences of horrific misuse of data within the sector, combined with anecdotal stories of (mis)treatment and usage of data having catastrophic effects within some of the world's most vulnerable communities, has highlighted for us the need for a book tackling issues of how we can all deal with data in a responsible and respectful way.

why this book?

What made 12 people descend upon a farmhouse in the Netherlands for three days, dedicating intense time and effort into creating this book from scratch?

It might have been an uneasy sense that the hype about a data revolution is overlooking both the rights of the people we're seeking to help and the potential for harm that accompanies data and technology in development context. The authors of this book believe that responsibility and ethics are integral to the handling of development data, and that as we continue to use data in new, powerful and

innovative ways, we have a moral obligation to do so responsibly and without causing or facilitating harm. At the same time, we are keenly aware that actually implementing responsible data practices involves navigating a very complex, and fast-evolving, minefield - one that most practititioners, fieldworkers, project designers and technologists have little expertise on. Yet.

We could have written another white paper that only we would read, or organised another conference that people would forget about. We tried instead to pool our collective expertise and concerns, to produce a practical guide that would help our peers and the wider development community to think through these issues. With the support of Hivos, Book Sprints and the engine room, this book was collaboratively produced (in the Bietenhaven farm, 40 minutes outside of Amsterdam) in just three days.

The team: Kristin Antin (engine room), Rory Byrne (Security First), Tin Geber (the engine room), Sacha van Geffen (Greenhost), Julia Hoffmann (Hivos), Malavika Jayaram (Berkman Center for Internet & Society, Harvard), Maliha Khan (Oxfam US), Tania Lee (International Rescue Committee), Zara Rahman (Open Knowledge), Crystal Simeoni (Hivos), Friedhelm Weinberg (Huridocs), Christopher Wilson (the engine room), facilitated by Barbara Rühling of Book Sprints.

about the title

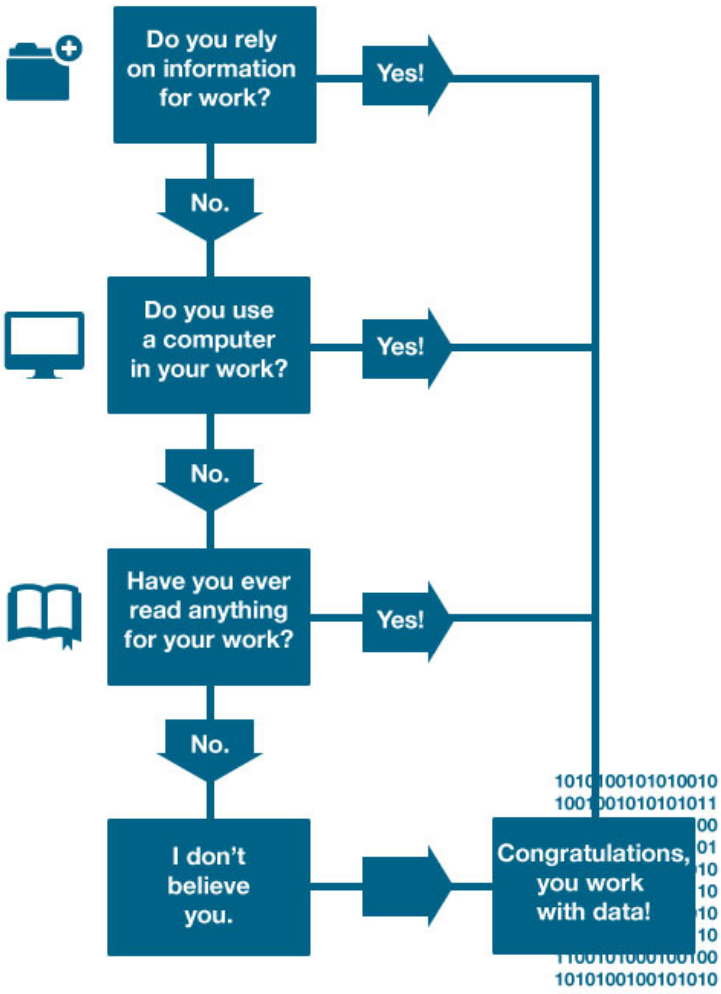
Shooting a harddrive into space is one way - though not a very environmentally sustainable way - to get rid of data. It's also an easy, yet seemingly impossible measure, whose consequences we don't fully understand. Responsible data isn't all that different, especially within international development. We operate in an information environment that is powerful and dynamic, and in which the hype and seemingly infinite potential of technology can easily distract us from what might go wrong and get lost.

Although we find it easy to deal with extreme cases and dramatic happenings, the nuanced ethical implications of how data changes our relationships with stakeholders and partners in country are harder to track. This book doesn't provide answers to these challenges, but we hope it might help. We'll have to work towards answers as we

continue to explore this new frontier. But hopefully it does provide some useful starting points, and point out a few exciting options (like the one in the title) that might not be the best responses.

am i working with data?

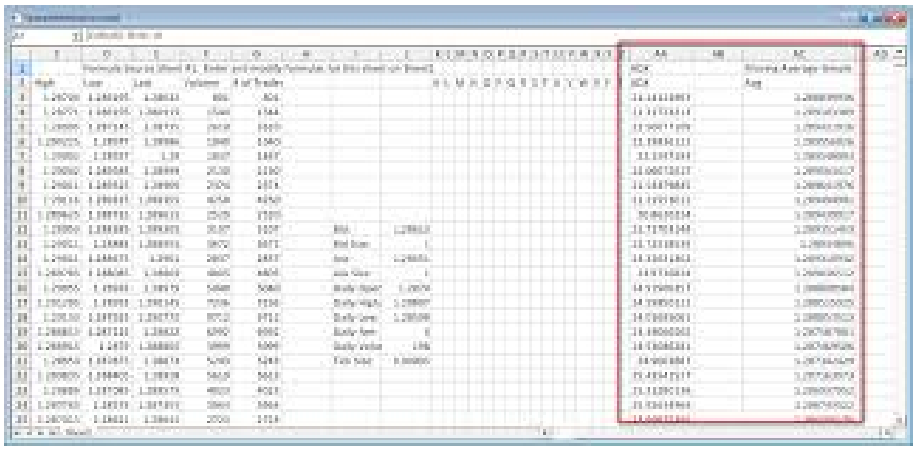
Yes, you are. Almost certainly. If information is in any way meaningful for the work you do, and if you plug anything into a power source, you are likely working with data.



what is data?

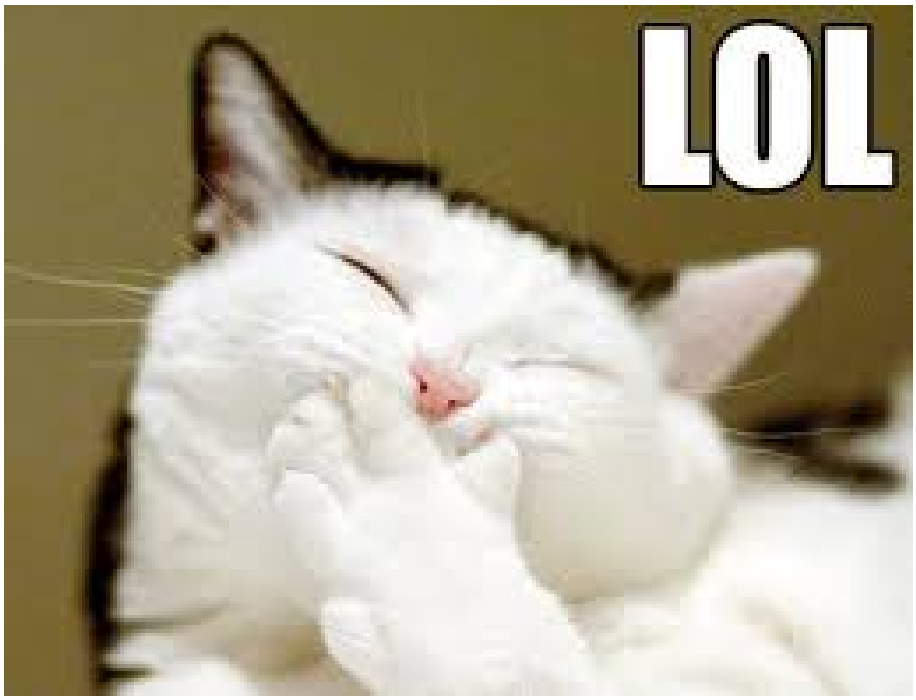
When we think of data, we often imagine large amounts of information that are formally collected, stored in a database or stashed in an excel sheet. Data is any kind of "value that is assigned to a thing", and can take a variety of forms.

So...this is data:



As is this:

But so is this:



...if you are studying LOL cats.

Pictures and videos online are perhaps the most obvious examples of new data, but we generate data all time. Writing emails, sending text mesages, automated GPS tracking on our phones. Data is everywhere, whether we recognize it or not. For this toolkit, data is any kind of information that is can be communicated and learned from, whether or not we call it data. See more at: http://schoolofdata.org/handbook/courses/what-is-data/.

who is in your data?

When thinking about responsible data, it's worth giving special attention to data on people. Data that reveals personal identities, habits, activities or affiliation is the most obvious point of care for responsible data practices. There are a number of data points that attach to individuals. At minimum, a person has a name, date of birth, weight, height, nationality etc. As more data is generated through a variety of mechanisms, however, there are all kinds of data that say something about who we are and how we behave, and often this data does not identify individuals in obvious ways.

But, data on people can include some unusual suspects; even data that doesn't have names could still potentially be related to people. In fact, even removing all personally identifiable information from a data set isn't necessarily enough to protect identities in that data set. With the increasing sophistication of analytical techniques and algorithms, "de-identified" data sets can be combined with other supposedly anonymous data to re-identify individuals and the data associated with them. This phenomenon, known as the Mosaic Effect, is particularly challenging because evaluating the risk of it occuring requires one to anticipate all the different types of data that exist or may be produced, and which could be combined with that data set, which simply isn't possible.

Even data that does not identify individuals can still reflect identities and behaviour of groups and communitities. Managing this data responsibly can be just as important as data about individuals, especially when data about those groups (where they reside, how they identify, what they do) might exacerbate social tensions, support discriminatory policy or incite violence.

Case study: wanting to act as an independent watchdog on the Election Commission of a country, an NGO set up an election monitoring platform, asking people who they were voting for. It was intended to be anonymous, but if that data set is accidentally made public and somehow combined with other data sets, it may well be possible to work out who voted for whom; this kind of information could be potentially very interesting to various political parties, and, dependent upon the political climate, be dangerous for the individuals in question.

whose responsibility is it anyway?

In short: yours.

Anyone producing, managing or sharing data that reflects on individuals has a responsibility to do so in a way that respects the rights and dignity of people reflected in the data, and avoids doing harm. Across a typical development project this will involve a number of team members working across the project life cycle, from planning to data collection, analysis, publication and outreach. Thinking about these issues can be difficult, and it's a reasonably new area of thought; to do this right, it requires a concerted effort by all team members, especially bringing together different disciplines and perspectives.

Project managers have a key role to play here, and official responsibility will always rest with the person (or entity) who can legally and organizationally control the data process. But ensuring appropriate and responsible processes will require the input and engagement of researchers, communications focal points, technical staff, enumerators and local consultants.

Managing data responsibly is **not** simply the task of one person; it's an overarching theme to be considered throughout (and after!) the lifecycle of a project.

the world of data: opportunities and risks

data is all around us...

This is the information age. Within international development (and beyond), data is becoming a necessary part of our work; thanks to technological advances, we can deliver, coordinate and communicate faster than ever. To put this into context, a full 90 percent of all the data in the world across history has been generated over the last two years. However, discussions and practices relating to the relevant ethical and political questions around data are not evolving at quite the same rate, and that is where this book comes in.

Some organisations working in the development sector have made data their main 'commodity' or 'service' - such as the UN Global Pulse Labs (http://www.unglobalpulse.org/) who analyse big data to accelerate social innovation for sustainable development; or Ushahidi (http://www.ushahidi.com/), who have developed a number of data management platforms to allow people to map out what is happening in their area. But 'data' is used by organisations in all sorts of ways - as information about whom your organisation is serving, statistics about the situation you're operating in, or even counting the number of countries in which your organization is active. [SEE section Am I working with data]

case study: refugee camp statistics

A humanitarian organisation is charged with delivering vital services to a refugee camp. In order to do so, they need to know how many people are living in the camp, and what their needs are. Also, it is important to note that once data is collected, it will not go away, and it can be used for purposes that you may never have thought of before. Rather than collecting information on people entering the camp on paper and by hand - a lengthy process, which is difficult to coordinate between the team - they now use a computerised system to keep basic statistics on who is in their camp. This allows them to get the right amount of supplies, reduces waste, and, if done correctly, will mean that all people entering the camp are included and receive the same service.

There's a tremendous amount of excitement about what data might imply for sustainable development. This is evidenced by the call for a 'global data revolution', and an increasing emphasis on measurement in response to some of the development sector's most intractible challenges. This excitement is importnat, and is stimulated by real opportunities, but has yet to include a critical debate about the potential harms that accompany increased use of data in development processes. Similarly, the buzz around 'evidence based decision making' has strenghtened interest in data and data collection among policy makers, without necessarily strengthening the capacities of data providers to operate responsibly or strenghten the capacities of data subjects.

walking the data tightrope : power dynamics

Data exercises power. It can create it, redistribute it, amplify it or disrupt it. It can entrench and privilege certain actors or perspectives, but it can also empower new voices and approaches. It can reveal and unravel atrocities, but it can also expose the vulnerable and marginalized. Responsible data ethics can often account for the difference between these binaries or polar extremes.

Collecting data about someone creates an inherent power imbalance to the extent that the data collector effectively owns a commodity relating intimately to an individual. This much is not new to those familiar with the "If it's free, you're the product" trope associated largely with social media networks. That data may be financially and practically valuable is widely accepted: that it could also have negative consequences for the data subject, such as whether or not they receive vital services, is less commonly understood. Similarly, we are familiar with the notion that data in the wrong hands can be dangerous, putting individuals at risk, but perhaps less aware that even in the "right" hands, there are violations that can arise from individuals being documented or categorized, leading to discrimination or exclusion. This could result in you paying more for a pair of jeans than a neighbor in a different zip code, or getting a higher interest rate because items in a shopping cart signal race or ethnicity.

This sort of “algorithmic bias”, widely prevalent in the marketing and advertising space, is increasingly finding its way into other kinds of decision making: about welfare benefits, immigration, healthcare and other sectors that you might be more

concerned about. Does this mean that you forego the opportunities that data presents and the societal benefits that it can facilitate? Sometimes, it just might. It may be that certain kinds of data are just too risky to collect, even as part of a human rights or development effort, for reasons that will be elaborated throughout this book. Or that none of the technical, legal or practical measures adopted to safeguard the data, and more importantly, the people that it relates to, really work. Those are extreme cases, however: usually, it is possible to manage the risks and achieve great successes and gains, while still being sensitive to asymmetries.

Issues of agency, legitimacy and representation pose additional problems in international development: the audience with which you are engaging may not be in a position to make informed choices or provide consent. Often, these communities are already lacking vital services and are unable to access their basic human rights; they might not be aware of the implications of their data being collected or used, have little or no awareness of their digital rights, and even less power to influence the process.

case study: iris scans

> The UNHCR are collecting biometric identification data (iris scans and fingerprints) from Syrian refugees who are living in Jordan, as lots of people arrive having lost their identity papers. They have shared the data with the Cairo-Amman bank, so that people can now get cash out from special ATMs by simply having their iris scanned, and have assured people that the data is staying 'just' between the UNHCR and the bank. But, if the Jordanian government were to ask for it, they would have to hand this data over.

These challenges place an even bigger responsibility in the hands of intermediaries, who need to be aware of the opportunities and risks of the data that they are working with, and embed sensitivity and responsibility in their data handling practices. However, it is all too easy, within development, to focus on the broader societal benefits that accrue, always for a 'good cause', and avoid the more problematic, critical discussion of how data is actually being managed or used, and whether they leave communities worse off after certain (arguably) paternalistic interventions than before.

We also see an increasing reliance on quantification: documenting, measuring, monitoring and reporting may be motivated by funders, by governments, by financial incentives or by research goals. This may lead to transformative or even incremental gains that hugely benefit people, but it may be at the expense of more nuanced, qualitative measures that do not override the legitimate rights of the less advantaged

who are impacted differently by the collection and use of such finely grained data. Adopting a critical approach to avert a “data for data’s sake” methodology will go a long way towards ensuring fairness and balance.

“Future-proofing” is also an issue: what seems unproblematic data right now, for example, may turn out to be very sensitive in the future. Changes to political situations or other ground realities may disproportionately impact certain communities relative to others. Having a long term view and working through various threat models can mitigate some of these risks. Minimizing the data collected can in itself be a responsible measure, regardless of circumstances.

Uncertainty is inevitable in these processes, and there is no perfect solution. There are good questions though, and asking these well in time, to the people involved, and to other subject matter experts, is always recommended. You may not be able to prevent every possible consequence of a data-intensive world, but being mindful of the particular vulnerabilities and circumstances of the worst off within the communities that you work for and with can go a long way towards averting or containing harm.

why responsible data

> Responsible data is*: "The duty to ensure people's rights to consent, privacy security and ownership around the information processes of collection, analysis, storage, presentation and reuse of data, while respecting the values of transparency and openness."*
> Responsible Data Forum, working definition, September 2014.

basic principles

Engaging with responsible data practices means upholding a certain set of ethical practices with regards to the way you use data. As the use of digital and mobile information becomes more commonplace in international development programming, we are only beginning to understand the implications that real-time information, data trails and information + communication technologies (ICTs) pose for relationships with and between the people and communities these projects aim to serve.

The power dynamics created by use of technology, or more specifically, of data, can complicate the ways in which we understand well-established norms like participation, consent, right to information and the freedoms of expression, association and privacy. Responsible data is a set of practices and considerations that aims to address these challenges in a practical sense, to help projects enhance the good they aim to do, and to avoid inadvertant harm.

At its base, understanding why responsible data is needed can be understood as a combination of **empowerment** and **avoidance of harm:**

Empowerment: We use data and technology to produce and mobilise appropriate information, to ensure that policies take everyone into account, especially focusing upon marginalised communities, that place the person in question as central to the data universe rather than a tangential bystander or even byproduct, empowering users to be active participants rather than passive data "subjects".

Harm avoidance: We do all we can to ensure that we 'do no harm' and that the way in which we use data and technology does not facilitate or exsacerbate harm done by others.

These norms are important for practitioners in the international development space because of our close and active engagement with some of the world's most vulnerable communities. As development programmes increasingly adopt innovative technical tools to achieve their objectives, the great potential of technology and data is accompanied by a professional, and moral, responsibility to protect the safety of those around us. Warranted or not, there is a level of moral responsibility held by organisations who claim to be helping the world's poorest people improve their lives. With great data comes great responsibility.

When we commit to managing data responsibly, we do so to improve our work and support of the work of those around us; to improve the lives of others; to avoid doing or enabling harm, and to actively pursue a positive effect on the world around us. Responsible data is not just about technical security and encryption, but about prioritising dignity, respect and privacy of the people we work with, and making sure that the people reflected in the data we use are counted and heard, and able to make informed decisions about their lives.

what could go wrong?

We don't know much about how data-driven projects can go wrong until they go terribly wrong. There are strong incentives not to share experiences of responsible data harm, and those who share stories, especially of dramatic harm, usually don't wish to be attributed. Nontheless, there are a number case studies described in this guide that illustrate the breadth of harm that can result from irresponsible data practices. Here are some broad examples of things that can go wrong:

Individals can be harmed physically, emotionally or financially. When personally identifiable information is leaked in sensitive contexts it can spark violence, discrimination, or exclusionary policies.

Groups can be harmed without individuals ever being identified, through the enactment of discriminatory policies on the basis of data, on the basis of percieved relationships or through subtle social dynamics or engineering.

Project credibility and relationships with local partners and beneficiaries can be harmed when stakeholders feel as though they are exploited for data without receiving benefits, or when projects have adverse and unintended consequences.

Organisations' brand and efficiency can be harmed, with negative consequences for funding, legal liability, high level policy discussions, or credibility with public institutions or the audience they seek to serve.

It takes just one tactical oversight, one data breach, or one mistakenly-collected dataset to put people in danger or damage critical relationships or organisational brand. You'll see throughout the book a number of case studies illustrating the scope of responsible data harm. Many of these are based on real-life experiences, and they are intended to highlight the sometimes frightening ease with which these situations arise. As with many of life's bad experiences - they can happen to any of us, at any time.

Whether the harm is dramatic and clear, such as violence and death, or more subtle and nuanced such as changes in social dynamics, this book aims to provide entry points and guidance for identifying and mitigating these risks before they occur.

arguments for speaking about responsible data with peers and management

Responsible data requires input and engagement across project teams and significant investments of time, energy and resources. This, in turn, will require getting many different types of people on board, and may require you to "sell" the idea up the management chain, convincing key decision makers of the value that can accrue and the necessity for proactive thinking and strategising on the topic.

This may require careful thought about different audiences, what drivers might engage them and motivate them to act, and what messaging might best filter through the organisation to embed these practices and approaches. Below are a few arguments that might be useful to begin discussions.

Developing a responsible data policy sends a clear signal to staff, donors and potential partner organisations about the organisation's progressive attitude and moral stance. Increasingly, donors are looking out for this, aware of increasing risks of security breaches and the reach of technology within the global development sector, and the growing ICT4D movement.

Taking clear steps towards responsible data practice demonstrates thought leadership within the sector and a clear awareness of the rapidly changing technological landscape in which we work. To be clear though, implementing such a policy within a large (I)NGO is no mean feat; it will require ongoing training, regular updating of the policy, and rethinking and reshaping of projects, but, as we have outlined already, it is without a doubt a necessary step. Whether you choose to take that step now or later is of course up to you; but the longer you leave it, the more people's wellbeings are put at risk.

Incorporating responsible data practices into project design is, simply, good project design. Addressing responsible data considerations when thinking about scope, audience, goals, risks, rewards and mitigation, will help projects to anticipate challenges before they arrive and plan accordingly, with positive consequences for the allocation of resources - human, financial and technical.

Responsible data practices make for more impactful programming. When done well, a responsible approach to data in programming can be hugely transformative and capable of disrupting or democratising existing structural inequalities. It can result both in internal benefits to organisations as well as systemic or environmental ones within the area you are operating in. It can minimize the extent and the nature of harm if/when something does go wrong, and it can contribute to a larger ecosystem founded not just on Do No Harm but on Do Good.

You may find that when discussing these issues with management and with colleaguest people assume that responsible data will come at a cost to some other important strategic objective. Generally, these tensions are worth discussing in groups. Working together to agree on the non-negotiable elements involved will generally help to identify responsible data practice, and is an important first step towards getting team members onboard.

perceived tensions you might come across

Transparency, openness and responsibility. There is a superficial contradiction between absolute transparency and absolute protection of individuals' privacy. At close look and in specific instances, this will often reveal itself to be a fundamental but manageable tension. When discussing specifics, it's rare to find a case where thoughtful, dedicated and informed people won't agree on what is permissable and what isn't permissable when promoting transparency and accountability. There is a key difference between personal data and data that should be made 'open'; as a broad rule, we believe that the right to privacy is for those without power, and transparency is for those with power.

Efficiency and responsibility. Often, objections to processes around responsible data will be hidden within objections to efficiency from the project implementer's perspective. Securing informed consent involves telling people about risks, and might decrease participation in a survey. De-identifying data may require external expertise and additional resources, delaying data release and delaying advocacy timelines. Humanitarian and other high pressure projects may feel that they should be exempt from all responsible constraints, since their work is, essentially, life and death. These tensions should also be addressed in the specific rather than the abstract, and teams should expect that they can agree on where to draw the red lines of what is absolutely not permissable and what benchmarks must be maintained. However, remember that respecting people's privacy rights is part and parcel of respecting their human rights; respecting these rights isn't an either/or, it's a must.

Representation and responsibility. Projects that aim to represent and amplify the voices of marginalised communities may struggle with the idea that some information about those groups should not be made public. The argument is sometimes made that it might be necessary to make some concessions in the pursuit of larger benefits; for example, causing some social discomfort or small harm in order to promote the position of a marginalised group, or empowering them to claim their rights.

further resources

- Donor code of conduct http://www.ssireview.org/blog/entry/a_new_donor_code_of_conduct
- UN data collection http://www.unglobalpulse.org/privacy-and-data-protection
- Professional standards for protection work carried out by humanitarian and human rights actors in armed conflict and other situations of violence: http://www.icrc.org/eng/resources/documents/publication/p0999.htm
- Fair Information Practice Principles http://www.nist.gov/nstic/NSTIC-FIPPs.pdf
- International Principles on the Application of Human Rights to Communications Surveillance https://en.necessaryandproportionate.org/
- OECD Privacy Principles http://oecdprivacy.org/
- UNFPA guidelines on data issues in Humanitarian Crisis situations https://www.unfpa.org/public/home/publications/pid/6253

misconceptions and common myths

If you are working with data and want to speak about the importance of responsible data, you may run into a number of recurring ideas that get into the way of moving this discussion forward - be it within your organization or in your interaction with other stakeholders, such as donors or beneficiaries of your projects.

Below are a number of those we have encountered so far and would like to address:

it is not *my* job, the it department has this covered.

Actually, IT staff don't always understand these challenges very well either; they likely have a whole other set of tasks to prioritise. And, no matter how intelligent or diligent they are, responsible data challenges take place in a no man's land where political processes meet digital information flows, where no one is an expert. The only way to responsibly address these challenges is to do so across expertise and project teams, and to think hard about every step of the data lifecycle - often in complicated information environments.

privacy isn't important to people who are struggling to survive, it's a western luxury that shouldn't be forced on us.

Everyone is entitled to basic human rights, and privacy rights (or digital rights) are just part of these. Responsible data is simply a way of understanding how best to respect the privacy rights of those around us. If the concept or the word 'privacy' is problematic, there are many other ways of framing it: respecting human rights, using information respectfully, being safe in the digital world. The consequences of violating someone's privacy rights could be just as serious as violating their other human rights - for example, leading to physical violence or preventing them from access to vital services.

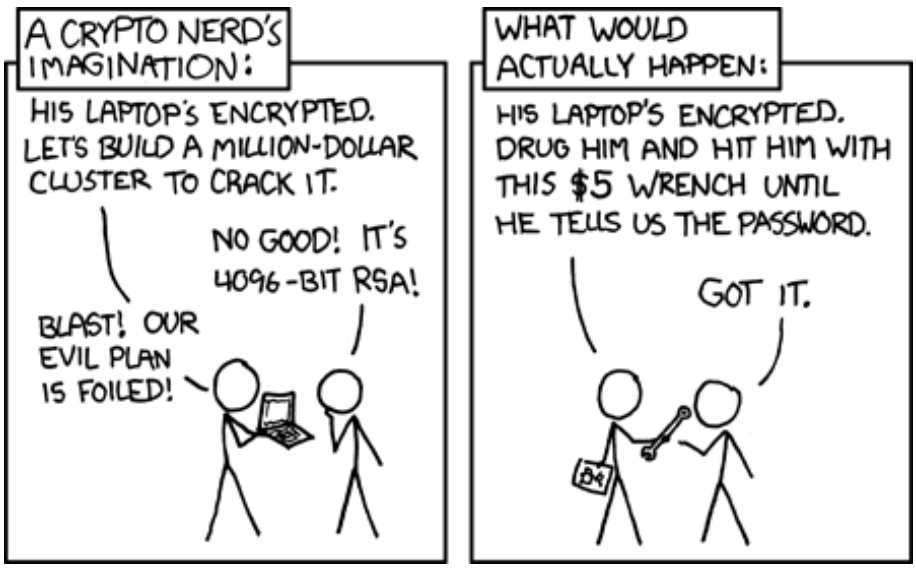
this is needless handwringing; no one has really ever been harmed by data.

False. There are actually a number of real life examples where irresponsible data practices have led to people being subjected to physical violence, in the extreme cases leading to multiple deaths, or discriminatory policies and infrastructure, and damaged credibility and relationships with projects. You'll read about some of them in this Toolkit.

all this hand wringing doesn't help, in fact it's distracting us from all the amazing things that data can do to improve livelihoods.

Responsible data isn't a zero sum game. Sometimes there are trade-offs to make, but it's generally possible to work within international development responsibly **and** without doing harm. It does require effort and investment though, and until organizations and individuals make a commitment to begin internalizing responsible data practices, there will likely continue to be accidental harm.

all our data is all encrypted, so it's fine.



Source: www.xkcd.com

There are some great technological tactics that we can use to improve our digital security - but it's rarely infallible. You could encrypt your hard drive in case someone steals it - but what if they steal it, then use physical threats to get the password?

it is all anonymised, so there's nothing to worry about.

Taking names out of data doesn't make it anonymous. There are multiple tactics for reidentifying data, and we have recently seen several instances where presumably anonymous data sets were combined with powerful algorithms to identify individuals and their online activity. This phenomenon, called the Mosaic effect, is especially problematic becuase it's not possible to estimate the chances that any given data set

can be re-identified, because it's not possible to anticipate all the data sets that might be produced and meshed with it. Let alone the impossibility of anticipating what kinds of data sets might be made available in the future...

this data is not sensitive.

Not sensitive for whom? Perhaps not for you, but have you asked the people who are reflected in the data whether they think it's sensitive? Will it be sensitive a year from now? What if the government changes, if political tensions are exacerbated or if discriminatory legislation is passed?

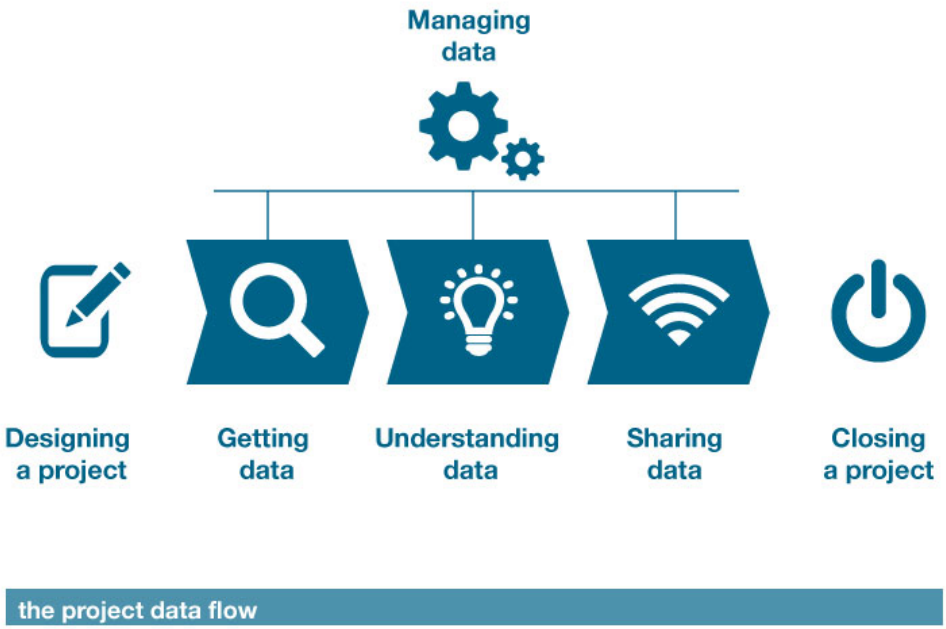
the people handling our data are absolutely trustworthy and have the best of intentions.

Even if that's true and will remain true in changing contexts, the road to hell is paved with good intentions. Responsible data challenges are complicated and require a thoughtful approach from multiple perspectives and expertise, not simply relying on the goodwill of a few individuals - human errors happen all the time, and it's best to have multiple strategies as a back up.

i don't have time for this right now; i have other priorities.

This, especially in international development, is almost always the case; we're working in high pressured environments. But just think what would happen if that dataset got leaked, or malicious actors got their hands on something they shouldn't... you could be putting the lives of the people you're trying to help, at serious risk.

how to use this book



the project data flow

Light Bulb designed by A.Dinagar from the Noun Project - www.thenounproject.com

We divided this book into thematic, independently readable sections that let you dig right into the part that is most relevant to your current objective. The six main sections try to cover the spectrum of considerations, practical methods and responsible data challenges in different phases of a project. While sections sometimes refer to other parts of the book, each module can be read independently. Should you prefer to read the book in the traditional way, the sections follow a narrative logic that is closely related to different steps of the project evolution.

designing a project

Thinking responsibly about data during a project is good, doing so before is even better. Designing a project to inherently accommodate responsible data practices works best. This section gives checklists, ideas and methods to inform your planning phase. It provides practical suggestions for thinking about risk assessment, threat modelling and budgeting to help you implement responsible data practices.

managing data

Any step of the project data flow will need to consider the management and storage of information. This section looks into how data is stored, where it is archived, who can access it and what the possible legal ramifications and risks when storing information are.

getting, understanding and sharing data

These sections get into the main pipeline of the project data flow. From zero to publish: what are the risks, perils, pitfalls, resources, tools and techniques you can use to develop a trusted, secure data flow that takes into account consent, agency, privacy and ethics.

closing a project

Harm can originate from zombie projects, abandoned databases, orphan sites and online ghost towns. This section is about getting closure: about secure disposal, archival and evaluation of your project.

designing a project

designing responsible data projects

don't panic, plan: assessing risks and threats

budgeting

designing responsible projects

Rapid changes in the way that information functions in development programming demands a careful consideration of responsible data challenges and practices. This requires engagement and input from all various expertise and perspectives across project teams, but will be the most efficient and impactful if addressed in project design. Here follow a few tips to help you get started on the right path.

baking 'responsible data' into your project design: where to start?

Project design can, if done well, translate seemingly 'fluffy' principles into tangible, concrete steps and activities. It can help narrow down a possibly bewildering array of tools and techniques into those best suited for achieving responsible outcomes, and set the scene for how you and your colleagues and partners will interact with data.

Here are some top tips to help you get your responsible design hat on.

- **Don't design alone:** Think carefully about who needs to be in the room, and engage the various stakeholders or experts as soon as possible (eg. it's usually a bad idea to bring specialists like techies or lawyers into planning processes after essential components have already been decided and can no longer be refined). Make sure that if you're asking people for their input, there is still space to incorporate said input into the project. Otherwise, you're wasting everyone's time.
- **Plan ahead:** map the lifecycle of the relevant project or system - it might change, but it's good to have a starting point.
- **Keep a timeline in mind:** Consider what a reasonable timespan might be- how long do you want to "futureproof" the practice or the system for? Be realistic here: it's impossible to know what's around the corner, but there are things you can do.
- **Factor in data:** Fill in the points when data may be relevant in this lifecycle and what the design implications are
- **Wait...why data?:** Think through the options for how data can be handled or processed at each point, and - importantly - consider alternatives to the obvious or usual way of doing things before deciding it is in fact the best approach
- **Get permission:** Consider what permissions, consents, policies or principles might dictate or affect the handling of data and the ability of end users to make informed decisions
- **Plan for failure:** Build in backstops, contingency and emergency plans for when things go horribly wrong (including a resource list for troubleshooting issues)

- **Budget what you need:** Ensure that the financial budgeting and the allocation of human resources are "fit for purpose" and adequate to achieve all of these objectives
- **Check yourself:** give yourself opportunities for monitoring and calibrating, and don't wait for the project to be shut down before learning lessons and parsing failures. Periodically check whether you are staying true to the responsible data standards and ethical framework that you set for yourself up front.
- **Document the process:** track both the small and big picture approach to have a clear, organizational baseline that can be relied upon as people come and go and the project evolves. This will also make it much easier for you to draw out lessons learned from the project.

don't panic, plan: assessing risks and threats

When working with data, there are many things that **can** go wrong, but that does not mean they actually **will**, or that the consequences will be catastrophic. Taking the time to do a risk assessment can save you a lot of time, and save you from future panics! It will help you identify what is likely and what is consequential, and thereby come up with a rational response. By carrying out a risk assessment, you'll preemptively come up with a number of preventative measures, as well as a back up plan to limit potential harms.

This section focuses on risk assessment practices specifically with regards to data, beyond digital and physical security risks, to include questions about the social impacts of data collection and publication, or how data can be reused towards objectionable ends. This does not, however, remove the need for a risk assessment to also be conducted for the entire environment of the project as part of standard planning.

threat modelling

Threat modelling is a type of risk assessment and part of a broader risk assessment framework. It is a useful process to help uncover specific threats to existing assets. To conduct a threat modelling exercise, discuss the below issues and identify how they manifest themselves for your project. (See the matrix template below these definitions).

assets

These are the types of data that the project will create or use (for examples of different types of data, see the introductory chapter, Are you Working With Data?). What data is being created? Where does the data exist and how do you interact with it? Map the types of data and the data flow. Does the data fall into clusters, such as public, internal, or confidential? If so, this already gives you a clue that different procedures may be needed for different types of data. (See section: The Sharing Spectrum)

risks

These are the vulnerabilities identified in the data flow, including access, sharing, storage and management of the various types of data.

adversaries

These are individuals and groups that may be interested in making the threats to your assets a reality. In other terms, they are the ones who you are up against. Research their **capabilities**, but don't let fear take over: when you know who they are, you will already have a clear idea of what you can do to stop malicious action. Also, do not forget that "adversaries" may be internal: the disgruntled former employee who goes rogue or the lazy systems administrator who forgets to do the backup.

threats

These are potential risks to your assets, to people or to the project. Threats to data assets can often be grouped into three generic categories: (a) loss, (b) illegitimate access, (c) manipulation. Threats to people can be sorted according to: (a) direct harm and (b) social effects. Threats to the project generally have to do with (a) sustainability and efficiency or (b) reputations and relationships. These generic categories can be made a lot more concrete for your given project, so use them as a guide, not an exhaustive list. Map out potential threats, paying specific attention to who or what you are working with. Remember that not all threats have adversaries or malicious intent.

likelihood

What are the chances that the above threats become real? Put differently, what is more probable to happen: an unidentified hacker who breaks into your highly secured systems only to prove a point, or you forgetting to budget for data hosting after the end of the project period?

mitigation strategy

Responsible data practices also require safety planning. This identifies actions you can take to address the threats. Questions that may help formulate your plan include:

> - What risks can be eliminated entirely and how?
> - Based on their likelihood and significance, which risks should be addressed first?
> - How can risks be reduced or better managed?

It is assumed that practitioners and managers won't be able to address all threats at once. They should be prepared to schedule work on project risk assessment and safety planning, alongside project design, implementation, and monitoring and evaluation activities.

The threat model can be organized into a matrix such as this:

Assets	Risks	Adversaries	Threats	Adversary Capacities	Likelihood	Mitigation Strategy

Writing these out, and thinking them through at the very start will help you to consider whether you have the resources you need to prevent the most probable risks, as well as the most consequential threats. It also helps you to work out your priorities and how you can meet the most urgent tasks, while being aware of disaster scenarios that might be catastrophic, but are less likely to occur.

Feed your findings back into your project lifecycle diagram. Does this leave any critical gaps? Do you need additional resources? Should you reconsider the project in its entirety? Asking these questions will help you focus on finding the right responses.

Clearly, you can't 'prioritise' or focus on every single scenario here: each organisation will have its own sense of whether it makes sense to focus on the more probable and less disastrous scenarios, or the less likely but potentially catastrophic ones. It may also help to develop a strategy for weighing different scenarios and allocating organizational resources accordingly. Either way, being aware of the spectrum of risks can help to arrive at a position and strategy that is tailored to your project's particular context.

BOX: Some risks are more avoidable than others. Data loss is a common and highly avoidable harm. Regular systems for back up are both accessible and affordable, and represents a type of basic data awareness that projects should review even in low-threat environments.

contingency planning

Prevention is only half of what you can do. There will always be residual risk and threats you cannot foresee. This is why it is important to prepare for incidents to happen, so you can contain them or mitigate their impact.

- Talk through the different kinds of risks, and make sure there is a clear, step-by-step plan of action to highlight incidents to the relevant people within the organisation if or when they happen.
- Encourage a culture of openness and learning from mistakes, not blame; reporting incidents within your project team should be seen as a positive, not a negative.
- Stress that the consequences of covering up or hiding in-house mistakes will always be much more serious than coming clean; mistakes happen, and it's good to learn from them.
- Keep a list of people to call in times of emergency in areas where you lack internal expertise - for example, lawyers, crisis communications, data forensics. Building those connections before you really need them will save you time if/when a crisis hits.
- For all projects, have an emergency plan of action to prepare for immediate shutdown or project termination.

external providers

Relying on external providers is often a necessity. It may actually be desirable to work with someone who is an expert in what they are doing, rather than spending your time on learning something that is hard to master and is outside of your core skills. However, ask yourself to what extent you can trust them, what mechanisms are available to you to verify this trust and what you need to know to evaluate what they are doing. If you're hiring external consultants to work directly with vulnerable populations to collect data, for example, make sure they come with strong references, or within your trusted network. Consider the option of independent audits or external references from organisations that you do trust, too, as well as the role of such providers over time, and establish safety mechanisms in case they become unavailable or their position or reliability changes.

holistic security

This is a way of thinking of all of the various security threats or risks that are faced: digital information security, physical and operational security, and psycho-social well being required for good security implementation. All three of these should be considered within technology projects, not just digital security, and there are many strong tools and support organisations who can help conduct security assessments - see Further Resources for a list of organisations working in this space. On the topic of holistic security, Tactical Technology Collective have put together a note with further discussion- https://tacticaltech.org/holistic-security.

budgeting

"do we have a line item for that? oops...."

Sometimes, treating data responsibly in your project will cost money and resources. If it's too late to include responsible data considerations in the budget, you will be in the awkward (and potentially dangerous) position of deciding whether not to be responsible; this could mean compromising on other goals, or coming up with makeshift approaches that may be unsustainable. Plan ahead for those budget lines that often are easily forgotten.

technology

Which technologies do you plan to use? Work out what you need before choosing a platform: be sure it has the functionality you need (or, if open source, that you can customise it accordingly) to grow along with your program and that you have budgeted for development, testing, maintenance and support. What will be the cost of using/replacing your technology in 1 year? 3 years? 5 years?

Does this technology meet your responsible data needs in terms of accessibility and security? Is the technology appropriate for the people who should be able to access it in terms of culture, literacy and media access? No technology is ever 100% secure, but technology that has been independently audited will have fewer vulnerabilities than non-audited counterparts. If your data is very sensitive, it may be more responsible to audit software before using it. This costs money, so ensure to budget for it.

support

Whenever you are relying on an external provider, especially for hosting or software, you will need support. There may be things that don't work, and there will be things that you only realise you will need after using a system for a longer time. Make sure that this support is covered in the agreement with your provider, or negotiate a support package included in the budget. Ensure that service levels and response times are appropriate for your way of working (eg. a 9-5pm solution for the UK may not be ideal if you are a real time crisis mapping organisation with a global footprint.)

Failing to think about these things in advance brings the risk that you will be running a system that you have considerable second-thoughts or grievances about, just because you cannot pay your provider more. As a business, they may well be unable to provide this support on a pro-bono basis, no matter how urgent or critical the need. Be sure to request non-standard packages that work for you; most providers offer non-profit rates and/or are able to support a fair price especially if their wider pool of customers can benefit from the tweaks or system improvements from your feedback and advice.

case study: vendor abandonment or planning for emergency data migration

An NGO has an internal policy of using free and open-source applications for data collection, mainly because of the accessibility factor (low/no cost). Many of this NGO's projects have invested in using this application to design their project workflows and have stored large amounts of data on the application server. The group who created the application decided to abandon the work for a new venture and the existing application was left unsupported. This resulted in the NGO losing access to their data for over a month during an emergency. Beneficiaries who had come to depend on the NGO's projects were left without service provision during a time of crisis. If the lack of vendor support was identified in the risk assessment, it's possible emergency/contingency planning could have allowed for a quicker recovery (e.g. emergency data migration) from this unfortunate situation.

hosting and data storage

Storing and hosting your data does not come for free and it is important to consider the cost over time in relation to how your programme may grow (look beyond year 1). Try not to take the inexpensive option simply because it is tempting to get hosting for 5 USD per month, although there may be instances where this is totally fine. Make sure your hosting is suitable for your needs, that is it is hosted with a provider you can trust or with competent systems administrators within your organisation. That usually means you will have a reliable backup system, regular monitoring of logs and swift updates of software to fix known security holes.

skills and training

Making sure that the people in your organisation have the skills they need is an essential, though often ignored, part of a project. This involves identifying the skills that need to be associated with specific people and specific roles, but also ensuring that no mission-critical skills are siloed with one particular 'expert' on a topic - what happens if they leave?

> - Do staff need security awareness training before setting out on data collection?
> - Do your project managers have adequate data literacy skills?
> - What do technical staff need to know about the communities you are working with, for example when designing the technical infrastructure you'll be using?

Addressing these kinds of questions when designing a project can help to enable appropriate responses to responsible data challenges, as well as generally ensure project sustainability and supporting staff's professional development.

Ongoing training can also be an important investment, to manage the risks posed by staff turnover and changing contexts.

backup planning and emergency support

Even with the very best planning, emergency situations may arise that will need immediate response and support. Though it's impossible to plan for everything, leaving some budget for these emergencies will enable your team to be more resilient. There is an almost never-ending range of emergency situations to consider for your project, of which examples might include:

> - recently changed threatening legislation
> - government crackdowns
> - violent acts and injury
> - imprisonment
> - loss of equipment
> - vendor abandonment

Responding to these unexpected challenges might require a many different kinds of resources, from external advocacy support, to emergency medical funds, litigation and legal support, increased security funds, or tool replacement funds. It is important to

remember that these situations are usually extremely time sensitive and turn-around time will need to be as quick as possible; financial bottlenecks should be reduced to a minimum.

quality and integrity assurance

Trust is important, but sometimes you will need to verify that best practices are being followed. At what stages in your project do you need to assure the quality of the work being done, in order to not let any irresponsible habits slip in? This can happen at the data collection stage (especially if you rely on interns, volunteers or short-term contractors), during data migrations and verification, or when data is analysed. Better safe than sorry, so factor in time and human resources to avoid unpleasant surprises when it is too late and the consequences are real, not theoretical.

project termination

Data-driven or data-supported projects beg one important question: what happens to the data when the project ends? Does the data have an expiry date? Where will it be archived? Might users ask for copies of their data? If so, are you in a position to do this, both financially and in terms of process? Make sure you plan in advance and have the appropriate budget to terminate the data, or to migrate it internally or to external partners as is appropriate.

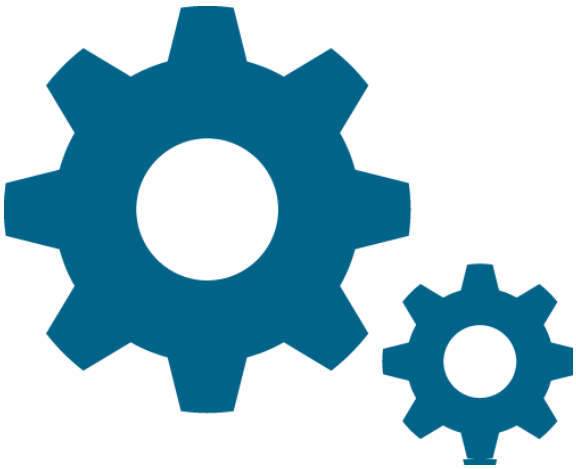
managing data

a home for healthy data

dude, where is my data?

for your eyes only...?

legal considerations



a home for healthy data

give your data a safe space

Data occupies physical space, even if in our imagination the bits and bytes are ephemeral. We store data on hard drives, and these drives have a physical presence, which brings a number of responsibility and privacy implications.

The most obvious issue to address is about data **access**: who can physically (or digitally) reach that hard drive and copy data from it, or simply take the drive away with them? Another one is **location**: there are many different places in which we can store our data. It can be on our laptop, on a USB stick, on a remote server in some other country, or even distributed throughout different geographic locations (interestingly, this is probably one of the most common cases). Location issues also bring with them **legal implications**, especially if we don't know in which country our data is being stored, and what laws govern digital property there.

This section provides an overview of approaches you can implement to ensure your information is stored, managed and accessed in a responsible, secure and protected manner. To begin with, it's worth reviewing the general principles of data integrity, and thinking about how to manage risks surrounding data storage.

data integrity

Data integrity is a term used do describe the validity, authenticity and security of information. It includes aspects of both information security and information quality, including how information systems and processes for sharing information are designed. As such, it has some overlap with confidentiality and availability, but is a useful frame of analysis for considering the interaction of technical and human influences on data responsibility.

- Is the data checked for changes at key points (upload, migration across servers, migration from data collection devices to the main server)?
- Are changes to the data flagged and are there mechanisms for version control?
- Are there granular permissions for making changes to the data during analysis?
- Is the data collection process well-documented? For example, through a codebook or other document that allows newcomers to the project to make sense of the data and understand the relationshipo between data structure, variables and the data collection process.
- Does the data include contextual information (potentially in *metadata*) that is necessary to understand the data?
- Are data collection notes (eg. on ambiguity, potential duplication, variations, or more contextual information such as notes made during personal interviews) provided together with the data in an accessible format, and the relationship between the two made clear?
- Are appropriate backup mechanisms in place in case of blackouts or system failures?
- Is the software used to manage and store the data fully up to date and licensed?
- Is the data interoperable? Can it be accessed by all the people and platforms that it needs to be? If people who might need to access the data don't have technical training, is it available in common file formats (such as csv files)?
- Have you future-proofed your data by thinking about how institutional and contextual factors might change and make the data difficult to access or use?

assessing and managing risks

A thorough discussion of risk assessments is included in the chapter on project design. When planning data storage, there are some specific risks and mitigation strategies that are worth considering.

Some data storage risks and harms:

- Loss of information (deliberate or accidental)
- Confiscation of information
- Data breach
- Legal threats
- Malicious attack

Some mitigation strategies:

- Speak to other organisations in your community who have conducted data projects successfully. Ask them about their storage practices.
- For bonus points, adapt or use standards for benchmarking your information security procedures, such as ISO 27001 or PCI DSS.
- Only store the minimal amount of data necessary to complete the task.
- De-identify data by default (though this approach has its limitations, see the chapter on anonymisation).
- Encrypt your data at all stages of its collection, usage, transmission and storage.
- Use secure tools (ideally open source tools which are more transparent) for communication (for a listing of secure tools, see https://www.prismbreak.org)
- Delete and destroy data safely when it has become insecure.
- Conduct regular audits and penetration tests of your security measures.
- Ensure your organisation is capable of managing and updating the system and tools on a long-term basis - as technology becomes outdated and easier to breach

For more information on protecting your information in a physical space, see Tactical Tech's Security in a Box chapter on this topic https://securityinabox.org/chapter-2

separate storage for sensitive data

You can store some information separately from the rest. For example, if you have a database with case files, you can replace the names with codes and create a spreadsheet with the codes and how they resolve to real names. You can then store this spreadsheet on another, highly-protected computer. This method alone will in most cases not lead to true anonymisation because the case files are easily linkable to real persons.

case study: hacked and then fined!

In 2014, the UK Information Commissioners Office (ICO) imposed a £200,000 fine on a charity called the British Pregnancy Advisory Service. The organisation's website had been hacked by an anti-abortion activist who threatened to publish the names, addresses, dates of birth and telephone numbers of using the service. The ICO determined that considering the sensitivity and risk of the information, the website did not have adequate security and left a vulnerability which could be exploited by the attacker. The organisation also breached the Data Protection Act by keeping data on it's callers for five years longer than was necessary.

Reference:

"Abortion service to appeal against £200,000 fine over hacked website"

http://www.theguardian.com/world/2014/mar/07/abortion-service-website-hacker-information-commissioner-fine

further resources

- The Frontline SMS Users' Guide to Data Integrity
http://www.frontlinesms.com/wp-content/uploads/2011/08/frontlinesms_userguide.pdf
- Deflect https://deflect.ca/
- Cloudflare https://www.cloudflare.com
- For a listing of secure tools, see https://www.prismbreak.org

dude, where's my data?

There are many ways to store data that vary in terms of convenience and security. How you store your data should be measured against the particular risks you or your partners may be facing, and your personal priorities: for example, are you more concerned about data loss, or surveillance? These are extremely important questions. Perhaps it isn't so important to you if authorities have access to your data, but it would be disastrous if the data were somehow destroyed.

As with lots of the points here though: no method of storing data is 100% safe, so it is **essential** that you backup your data no matter how or where it is stored.

data storage

There are a few areas to take into consideration as you make decisions on storage and these could include:

- **Physical location:** Where should data be stored given the type of data and the potential vulnerabilities? Within the country you're working in, or in another country, and what are the implications of both of these? (eg. differing legal jurisdictions, local internet regulations, especially with regards to storing sensitive data)
- **Digital location:** Should data be stored on-line or off-line or both? Or is it open data? (see: The Sharing Spectrum)
- **Ultimate ownership of storage:** Consider pros/cons of third party data storage vs local data storage (e.g. do you have local capacity for local data storage; what is the desired uptime; etc.)
- **Back ups:** what level of data back-ups are required? (you can never back up too much!)
- **Access:** ensure storage method allows for the necessary levels of access. (see: For Your Eyes Only)
- **Data lifecycle:** what is the data lifecycle and who will manage it until the end of the data's life? (if it ends!)
- **Saving data:** what is the data access / data saving policy, especially for sensitive data (e.g. if data must be downloaded locally it must be encrypted, etc.)

pros and cons of storage options

Some of the advantages and disadvantages of various approaches may be:

locally on your pc(s)

- **Advantages:** high security if encrypted, fast and always accessible, easier version control, greater legal clarity over ownership
- **Disadvantages:** lower security if physical theft or confiscation is an issue, poor backup, bad resiliency in terms of physical emergencies - fires or floods for example, poor access except for people located near the machines, potential increase in time spent maintaining the system versus working with a specialist provider.

on your own network

Advantages: increased resilience, easier sharing and collaboration, better backup, greater legal clarity

Disadvantages: increased cost, greater skill needed for effective security, greater reliance on IT support team or third party contractor, potential exposure still to confiscation

living in the cloud

- **Advantages:** decreased costs, increased resilience, easier sharing and collaboration, less downtime, better backup, specialist provider knowledge
- **Disadvantages:** increased legal uncertainty, no control over physical access, nearly impossible to "vet" people with access to the servers, reliant on "cloud" company policies which may change

If using "cloud" hosts, you should ideally identify those who have previous experience in:

- Providing secure access and protection for your information
- Supporting the tools which you wish to use
- Providing good and fast support to their customers - find people who come recommended from trusted members of your network
- Demonstrable and veriable commitment to privacy
- Working with other projects in the humanitarian or human rights field
- Working within the legal jurisdiction you need

protecting physical data: securing your information from theft, damage and loss

For many people, information security often makes them think only of digital data. However, physical data protection is a vital process, as gaining physical access to data often requires less technical skill than a cyber threat and can often be an easier strategy for a potential adversary.

Imagine if your office or home were burned down or broken into today - what would you wish you had thought of in advance? Some things to think about include the following:

- **Locations of data:** the physical security of locations where you store physical data (such as paper) or physical media (such as laptops, USB sticks, DVDs, SD Cards, hard-drives). Are they all in the same location, and easy to find and gain access to?
- **Access to sensitive locations:** who has access to your office, home and working environment. Especially to areas of highest sensitivity such as server rooms, research desks, consultation rooms, meeting rooms etc. Consider using high-grade locks, CCTV, fences etc.
- Installing appropriate fire safety controls
- **External staff:** vetting all staff and contractors such as cleaners or security guards
- **Using an inventory:** creating and consistently updating an inventory, to enable you to be able to identify any loss or theft of data.
- **Building a security incident registry:** This should be filled out if any member of staff physically sees anything suspicious or has an unusual incident occurring with their IT equipment. It allows for monitoring and identification of incident patterns which may otherwise have been missed.
    - For help in identifying suspicious digital incidents, see http://digitaldefenders.org/digitalfirstaid/

- **Getting rid of physical waste:** regularly shredding and desposing of any paper waste

- **Regularly changing security procedures:** for example, changing keys, cards, pincodes or other access control mechanisms, Particularly following a change of staff.

further resources

- On choosing a hosting provider, see https://learn.equalit.ie/wiki/Responsible_Data_Forum_on_Hosting
- On setting up a secure hosting provider, see https://learn.equalit.ie/wiki/Secure_hosting_guide

for your eyes only...?

who can access the data?

The person ultimately responsible for the data will need to have both physical and digital control over the data. Physical access control means having control over the documents, computers, servers and working areas where data is kept. On the digital side, check that are you doing all you reasonably can to secure your data - for example, using strong passwords, encrypted connectons, VPN, logging, two factor authentication etc.

To work out what kinds of permissions and controls others in your team might need, mapping the types of users and what kind of access they will need is a useful exercise to undertake early in project planning.

case study: "human trojans"

Many NGOs work on areas that involve high stakes and powerful interests, such as when it comes to operations in natural resource issues (oil, diamonds, gas). This often involves collecting information 'on the ground' in countries where those resource are extracted (from governments, companies, local people) and then transferring this information to an international office in another country – often in the Global North – for analysis and reporting. In this case, it was a company that was adversely affected by one of the NGO's reports that started wondering where the leaks were that caused negative information on their practices to end up in the NGO reports. It then hired private intelligence services to specifically find ways to get people working at the international offices on their payroll. This time, the attempt to infiltrate the international office to find out what the sources where coming from was unsuccessful since some of the targeted employee reported it. The assumed intention of the attempt was said to be to find out embarrassing information on the NGO to use as leverage, potentially embarrass their donors, bug their offices and place trojans on their computers.

Lessons:

- Human beings are the biggest risk
- Provide access to information on a need-to-know basis only

> • Even if local operations are secure, complacency in international offices may increase vulnerability

> Mitigation:
> • Strong vetting of staff is crucial
> • Conduct and continuously update a threat assessment, especially when new data is released
> • Understand your potential adversary
> • Think also about physical access control of offices

setting permissions

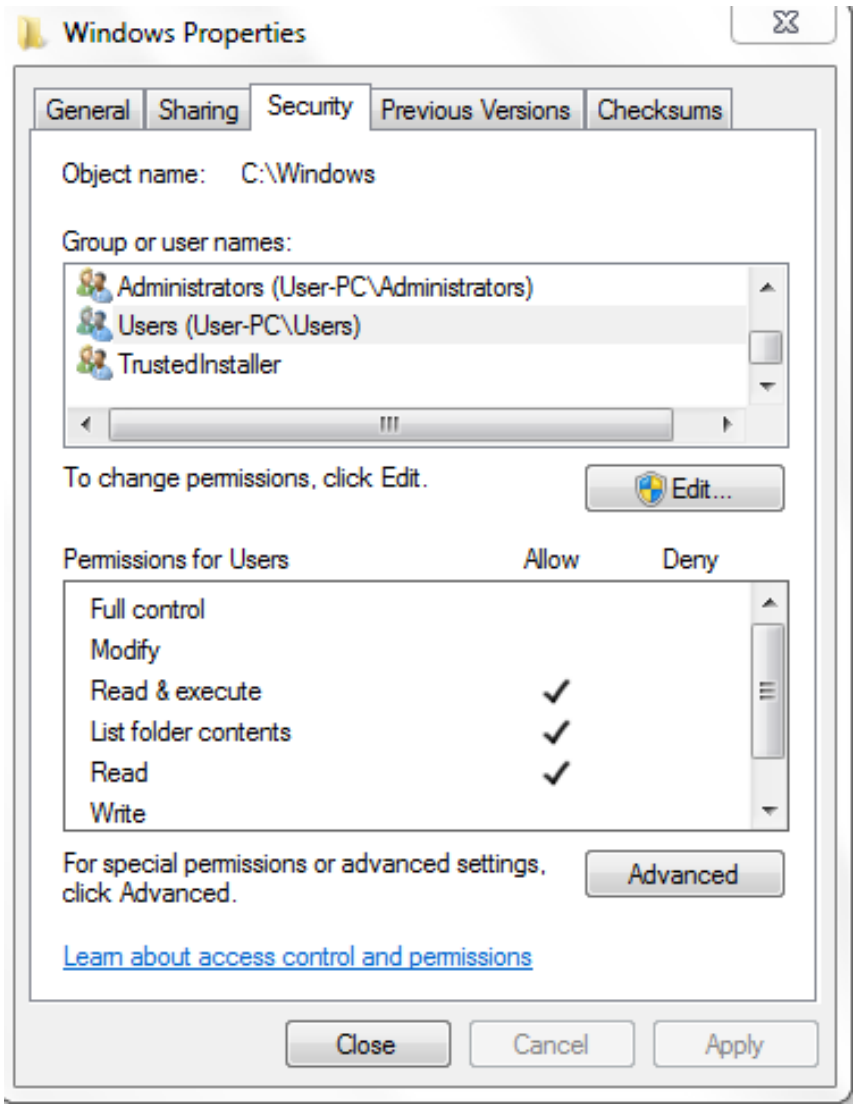
A general security principal is to **limit user access to only the minimum amount of data** that they need to be able to do their work effectively - and **only have access for as long as they need it**. This ensures the privacy of people who have had their data stored by your organisation, and also reduces the potential impact of security incidents such as a data breach or data loss. Put otherwise: if only very few people have access to the whole database, then there is less chance of them accidentally deleting all of it.

People should also **only have the minimum permissions that they need to do their job**. To do this, you must think about what the broad categories of users will be and what information they need to work with. For example, a role such as a remote medical clinic worker on a short term project might only need to access data relevant to the area they are travelling in for a few weeks of their deployment. This means they only need to have access to office files and IT systems which are relevant to that specific area. However, someone working as a country researcher tasked with identifying trends across the country might need access permissions to all of the data in the country for a longer period. Although it might be less work for you to simply give out the widest possible permissions to people, to avoid them having to come back to you and ask to adjust them - beware! The easy route is not always the most sensible, and doing so could create a lot more work for you in the future.

Access permission also includes items like auditing and controlling who can manipulate, manage and delete data. The person with responsibility for managing and creating access should schedule regular times to audit the users types and remove people who no longer need access - for example, temporary workers such as

contractors and interns. Unfortunately this is particularly important for potentially digruntled staff such as workers who are fired - a number of examples exist of people using their old job credentials to illegally access data at their previous roles.

Example of access control and permissions:



HUMAN RESOURCES: Bear in mind that if you are introducing people to a new set of tools or methods for accessing the data, they will likely require ongoing training sessions and follow up.

case study: "donor data danger"

In a project involving sensitive human rights work, two of the donors involved insisted collecting a huge amount of data on operations, including names of people involved and receipts for all activities. Within three months of the project, however, both of those donors discovered inside threats. In each case, some of their employees had become disgruntled with their working conditions and subsequently left the organisations, taking with them a large amount of data – not only on one project, but all the NGOs and individuals the donor had been working with and/or funded. In some instances, this information was even chanelled to the country's intelligence services directly or indirectly, exposing a large amount of people to large risk (including employees, sources, beneficiaries, as well as their connections and networks). At the high point of this crisis, even the regional director of one of those donors said we should stop passing sensitive information to his own employees, as he couldn't trust those in the organisation. While it is impossible to prove causality, a considerable increase in attacks and disruption of those NGOs' work that had linkages with those particular donors was reported.

Lessons:

- Be aware and prepared to actively disagree with your stakeholders, even if they are your funders
- International actors can be just as much a threat as anybody else – secure your info from the bottom up, but also all the way up
- Don't think because an organisation has an international reputation, they are automatically responsible data holders – in the end, you will be responsible for your data and the people it reflects!
- Encourage a culture of responsible access controls and permissions not just within your organisation, but also in organisations you share your data with.

collaboration

Controlling access effectively also means selecting the right choice of methods and tools which balance the need to keep information secure and also allowing effective collaboration in the field. A number of things which need to be considered in this area include:

- **Types of data** that people are collecting, and where they are putting it.
- **Tool choice:** suitability, useability, support, updates, cost, local vs network, network vs online access, mobile vs Desktop, open source vs proprietry.
- **Levels of verification for the data:** how many people should be looking at and checking the data?
- **Speed of internet connections for users accessing data:** eg. if users are collecting data on their mobile devices in the field, the mechanism chosen for collaboration will need to be able to cope with receiving from such devices. Depending on phone signal, this may mean collecting and sending video from remote locations is not possible. This should be thought about **prior** to committing to any technical infrastructure.
- **Access points:** Collaboration often requires remote access which can occasionally decrease security, as it opens up a number of less easily secured access points to a network. As such, tools and methods such as forcing regular password changes, two factor authentication, network logging, VPN only access, etc. are important to help mitigate such risks. (see: Security Resources)
- **Layered access:** Having appropriate access permissions is pivotal to ensuring that strategies on separating information work. Note that it can also be an option to only allow access when several people co-sign, that is, certain data is only accessible when more than one person unlocks it.

but don't overdo it!

It's important to make sure that there are rigorous controls on how data is accessed, but you don't want to lock it up so tight that you can't get the data when you ned. In fact, projects should take active steps to make sure that the data can be accessed by the right people at the right times. Below are some concrete steps to take that will ensure data integrity, while also making sure that it's available only for the right people at the right time.

- Backing up your data: regularly back up your files, and test these procedures on a regular basis.
- Building **redundant systems** in case of failure: For example, it can happen that hard-drives or essential equipment can completely fail without warning. Without redundancy like alternate data servers, this can completely disrupt your operations.

- Building architecture with extra capacity: data collection can be disrupted when space runs out on storage or collection platforms. You should try to ensure any system you have can be easily and cheaply expandible.
- Emergency situations: unexpected interruptions like natural disasters or internet shut dows can interupt data collection and transmition. Having a plan for such occurances (for example, to shift to satellite links) can minimize the damage they do.
- Targeted disruption: sometimes someone, or something, may target your data specifically. Ideally, threats such as this will be identified in a risk assessment, allowing you to build extra capacity into your architecture to deal with such problems. If hosting on your own, fighting against malicious attacks (such as Deliberate Denial of Service) can be very expensive. However, services exist which specialise in absorbing such attacks and are recommend good practice - such as Deflect (https://deflect.ca/) and the commercial provider Cloudflare (https://www.cloudflare.com)

legal considerations

All steps of the data lifecycle are subject to legal requirements, and managing data securely requires understanding and meeeting these requirements adequately. It's better to be proactively aware of the legal restrictions on your work, than to realise after the fact that you've been breaking the law and face monetary fines as a consequence, or that you can't legally do what you were planning to do. With the growth of cheap *cloud computing* (ie. data stored on servers not owned by your organisation), it is often difficult to know exactly where your digital data is being held. Popular providers of email and storage such as Google, Amazon, Facebook, Yahoo, DropBox, GitHub make use of infrastructure spread across a number of international sites. For example, Google stores data in the US, Ireland, Belgium, Finland, Chile, Taiwan and Singapore; so, which country's laws affect your data? It's not always easy, but you need to understand the legal ramifications of where you store your data.

what types of laws and procedures apply to your data project?

jurisdiction

It can be challenging to understand what countries' laws govern the management of your data. Some common grounds for jurisdiction include:

> - The countries where your organization is registered
> - The countries where your organization operates
> - The countries where your data is stored (for cloud storage, this can quickly become complicated and involve multiple countries)
> - The countries where your users, participants or subjects are

Sometimes the terms of service or specific policies will determine which law applies, but often not, and other jurisdictional claims can supercede these. As a point of departure, it's worth assuming that all of these jurisdictions apply. Talk to your technical and legal team to determine which don't.

data protection laws

A number of countries have strong data protection laws which place limits on the types of data which may be collected from individuals. They also often include specific legal requirements about the methods used to store such data, along with mandatory reporting and monetary fines for breaches. Individuals about whom data is stored are often granted a number of rights, such as access to their information and the right to have their data correction and/or removed.

For an overview of data protection laws in different countries, you can browse http://www.forrestertools.com/heatmap/ and http://www.dlapiperdataprotection.com/#handbook/world-map-section.

It might require some careful thinking, but granting people rights to access data in which they are reflected is crucial, especially for organisations which collect sensitive data, such as witnesses of human rights violations and perperators.

case study: data protection laws

Organization X was a human rights organisation in Sub-Saharan Africa collecting and publishing data about human rights abuses by the local government. The government was embarrassed by this and wanted to stop the organisation from functioning effectively. Rather than attack the organisation physically, which they knew would draw international attention, they decided to disrupt its work by tying them up in nefarious legal cases. For example, the organsation was severly punished for poorly protecting data, minor health and safety violations and accounting malpractices. While the cases were all eventually thrown out, this caused harmful disruptions to the organisation's work and caused them to spend their limited resources on lawyers' fees. Adhering to relavent data protection laws may not prevent this kind of legal tactic to be used against you, especially if the law-makers are targeting you specifically, but it may limit their techniques.

encryption technology laws

Local laws in a number of countries (such as Sudan, Yemen and Pakistan) place limits upon the nature of encryption software allowed for the communication and storage of data. System architecture and tool choice must incorporate these concerns. Other laws which effect the use of encryption include:

- mandatory handing over of encrypted data if requested by government authorities
- mandatory metadata collection by specific industries, eg. telecommunications (and potentially their subsequent sharing with government authorities)
- laws requiring the weakening of encryption software, such as programmes for export: some countries (eg. Pakistan) won't allow the use of programmes which contain certain levels of encryption. Other countries (like the US) try to control the distribution of programmes with high-strength cryptographic controls - this originated from countries not wanting to share their high-strength cryptographic software with potential adversaries.
- those which require individuals to disclose their passwords (such as the UK) upon government request.

ngo registration laws

An increasing trend in many countries has been the introduction of strong laws which regulate the presence, funding and/or activities of NGOs in their countries (for example, Russia, Ethiopia, Egypt, Hungary, Kenya and South Sudan). Projects initiated in the country must consider these when building data management infrastructure, and when thinking about different country presences.

jurisdictional issues

Organisations should be aware of laws that would give governments access to information stored on servers in particular countries. For example, Boston College was forced to give interview information (tapes) to the Police Service of Northern Ireland after they were subpoenaed. Cite: http://www.timeshighereducation.co.uk/features/oral-history-where-next-after-the-belfast-project/2013679.article

It is very difficult to protect digital information from subpoenas (for example, see this map on US extradition treaties http://qz.com/97428/map-how-to-stay-out-of-reach-of-us-extradition-treaties/) so it is important to adhere to a minimalist approach to collecting or storing sensitive digital data (or don't store it all) - see section on Getting Data.

Organisations must also be aware of cross-jurisdictional issues in relation to their data management. It is not unusual for data to be collected in a country office, transferred to a regional office in another country, then onwards to the organisation's

headquarters in a third country. Also, some states or regional groups place strict conditions on where their citizens' data may be transfered to - for example, the EU/US Safe Harbour law.

copyright and patent

Copyright and patent issues related to the collection, storage and dissemination of your data are important laws to consider. Ensuring you have the correct licensing for your projects and infrastucture can help reduce unexpected restrictions and costs further into your project. Open licencing such as Creative Commons options, and open source code such as MIT or BSD licences provide tested and off-the-shelf solutions for your data, and open the door to your peers being able to verify your data, or mixing it with other datasets that you might not have collected yourself, but which would strengthen your project. (See section: Disseminating Data for more information on licensing and advantages of open licensing)

procedures on presentation of evidence

If you intend on using your data in legal cases, you should understand the requirements used in court for the presenting of evidence. Approaches to digital data differ significantly in each jurisdiction. However, some standard basic requirements are that data management systems:

- Allow for an auditable chain of custody (ie. make it possible to know who has had access to the data at all times)
- Ensure the integrity and authenticity of the data. For example, a number of technical processes (such as hashing and forensic examination) allow a comparison between original and other versions. Authenticity can be enhanced through the collection of extra information such as *metadata*. However this must be balanced against the extra risk this may pose to collectors - as often it is not clearly understood by the people involved.
- Ensure that data are verifiable. The data content itself in many cases should be verifiable. For example, the massive growth of online video posted to sites such as Youtube, Vimeo and LiveLeak has driven the development of methodologies to verify what is shown in the footage - is it showing what it says it is showing?

further resources

- NGO Law Monitor - http://www.icnl.org/research/monitor/
- Maps of Data Protection Laws - http://www.forrestertools.com/heatmap/ & http://www.dlapiperdataprotection.com/#handbook/world-map-section
- Choosing an Open Source Licence for Code - http://choosealicense.com
- Creative Commons - https://creativecommons.org

getting data

what's your question?

collecting new data

working with existing data

power to the people

consent

what's your question?

Any evidence-based project relies on the complex, nuanced and deeply important process of collecting data. This chapter will address considerations around data collection: from defining your question, through to practical procedures for dealing with new and found data, to questions of agency, consent and privacy.

Determining the questions that you are trying to answer is the most important part of a data collection process. Data should always be gathered for a reason: namely, because you think the answer to your question lies within. You may be asking simple **factual** questions (eg. how many students graduate from a school annually), or questions that are **divergent** or **evaluative** (what impact did the change in curriculum have on students graduating from a school). No matter how simple or how complex, **knowing the question that you are trying to answer is fundamental**.

The next step is to determine where to find information that will most effectively help you answer your question. There are two main ways of getting hold of data:

Using existing data – data that has already been released and is publicly available, like a government census. This group also covers accessing non-released existing data (like through *Freedom of Information requests*), or taking data from existing sources that

hasn't been clearly packaged for research purposes, like scraping social media sources like Twitter and Facebook.

Collecting new data – gathering data directly from the subject (e.g. doing a survey of students and teachers in a school) or generating new data,
potentially indirectly (e.g. logging GPS pings of staff cell phones while on specific missions).

No matter where data come from, **try your best to ensure that it is good data**. While there are many things to keep in mind about data quality, the most basic ones are **accuracy**, **validity** and **timeliness**. These are extremely important aspects of using data responsibly, but the technical sides to these, for example on data collection techniques and methodgy, fall outside the scope of this book. Never fear though - there are many sources on this issue that can be referred to for help.

further resources

- Data QualYtI: Do you trust your data? (an article Hjusein Tjurkmen, Mariyana Hristova, Musala Soft) http://istabg.org/data-qualyti-do-you-trust-your-data
- See more at: http://schoolofdata.org/handbook/courses/finding-data/#sthash.8o6YozUJ.dpuf

collecting new data

designing a data collection process

Designing the data collection process requires asking basic questions about roles, tools and timing. This section needs to be informed by the **risk assessment** conducted during project design (see Don't panic, plan.), but reviewed in detail here for your own project, to carefully consider responsible data challenges that might surface by the particulars of your data collection. To begin understanding what data collection methods will responsibly produce the data required by your project, **review the following questions**:

- **What** are you collecting and **why** are you collecting it? How does this relate to your research question? Have this clear from the beginning of your design as it helps you throughout the process in making the right choices and changes.
- **Who provides the data?** Did you get their **consent** for this specific use?
- **Who will collect** the information? Ensure they are appropriate in relation to the specific identities of your participants (e.g. gender, race, class, language, etc.)
- **How will you collect the data?** Establish a clear data collection process – will data collectors perform data entry in front of responders or after a conversation with responders? Consider the impact of the data collection method on the desired conversational outcomes between the data collector and respondents (e.g. conversations about sensitive topics shouldn't be interrupted by the data collection tool; do what you can to remove external factors that might act as inhibitors for respondents).
- **How often** will the data be collected? Will the same participants be polled repeatedly? How will frequency interact with other timeframes important to the community in question? What implications will this have for project resources and for data validity?
- Take careful consideration of the **tools** you will use to perform data collection. In some cases, paper is a good option if using electronic devices will create suspicion or risk. But also consider that in some cases, electronic devices can be password protected and encrypted to protect the data in ways that paper cannot.

- How will you **transfer and store** the data? Will the the data collected need to be centralised? In some cases, decentralised data is appropriate because the data is mainly operational and serves a short-lived purpose. If you plan to centralise the data, consider the data flow process from data collection to centralisation. What is the desired frequency of data centralisation? How secure is the data centralisation process, and what additional risks may arise from centralisation?

defining challenge areas - an exercise

step 1: come up with a detailed data collection plan

The questions above can be used to define specific challenge areas for your project. You might want to **create a table** to map your steps, adding specific approaches and challenges that might arise for each. It would be best if done in a group, on a big sheet of paper. The table below shows examples of what specific approaches might involve.

DIMENSION	SPECIFIC APPROACH
FROM WHOM	All adults
HOW	Household surveys, conducted in the local language
WHO	local enumerators, hired through the national university
HOW OFTEN	annually

step 2: brainstorm risks and responses

When you have identified some of the specifics, talk about some of the common responsible data challenge areas, such as identification and the social impact of data collection. Do any of these prompt concerns about the methods that you are planning to use to collect data? Write down any potential risks that occur to the team during discussion and note any potential responses that might help to manage those risks.

CHALLENGE AREAS	POTENTIAL RISKS	POTENTIAL RESPONSES
Personal privacy and anonymity		
Agency and data empowerment		
Social Impact of data collection		
...		
...		

deciding what data to collect

data minimums and unanticipated use

When deciding what data to collect, consider the minimum amount of data necessary to reach the project goals in balance with the data you may want to collect for unanticipated uses. While many development programmes track clearly defined indicators, there may be an impulse to collect additional data outside of indicator tracking for operational or research purposes. It is important to refer back to your risk assessment to inform these decisions.

red lines - what not to collect, and what never to collect

Be clear about what information you will not collect due to relevance, sensitivities, and risk. Think about unanticipated uses of the data and potential harm/risk.

In general, when thinking about what data to collect or not to collect, these are the most important aspects to consider:

Leave it out: Do you really need the names, ages and exact locations? The rationale for collection should be that it **needs to be necessary to achieve your goals;** not just that it would make things easier for you, or save you time. Often you can reconcile this by collecting less fine-grained data. Sometimes it means you will decide to refrain from collecting some data completely.

Metadata: This is data describing your data, for example the time a photo has been taken or the camera that was used. This metadata can reveal a lot of the information that you were deliberately not collecting, or that you want to protect. If this is the case, make purging of metadata part of the collection process from the very beginning, or configure the tools you are using to not create it in the first place.

Surveillance: While you are collecting data, you are not always alone, and when someone is passively or actively watching you, they can fill in the gaps. Depending on your risk assessment, this means you need to take appropriate measures to conceal your traces in the collection process, alter collection methods, or not collect the data at all.

live data processes

Some projects will use "live data", which is collected, accessed, reviewed, used and updated on a continual basis. Examples of such projects include service delivery projects, or projects that are case based, in which several different team members have access to individuals' data and records. These systems can be especially challenging, as they compress the project data cycle, and force individual team members to adopt a variety of roles and responsibilities when relating to potentially sensitive data.

roles and responsibilities

The first step is to clearly define **who should have access to exactly what data at which point.** For example, during the registration process, you may have intake staff (consider maybe these staff are volunteers or interns) who should be able to create new registration records but should not have access to detailed assessment information (which may contain sensitive information) about the subject. In this example, it could be that only case workers should have access to detailed assessment information on the subject. If referring a subject to another agency, that agency may receive a specific sub-set of information on the subject as to protect sensitive or non-relevant information.

TOOLS: Create a Roles-Responsibilities-Functionality Map to help you identify who should do what and when. Here is an example:

	Data Collector	Data Analyst	Data Collection Supervisor	Program Officer (Local Office)	Program Director (Central Office)	Donor
Build survey form						
Data entry on Mobile Device						
Data entry in Browser						
Manage data sync A						
Manage data sync B						
Export data						
Internal Reports						
Dashboard						

Legend: Create/Edit; View Only; No Access

access to data throughout the *micro-cycle*

In addition to defining roles and permissions during the service delivery period, you should also define roles and responsibilities throughout the project lifecycle. Consider who will handle data (and who should not) during data collection, data management (storing, cleaning, analysing), and internal data sharing. Do you have data administrators who will oversee the data workflow and support data access and flow according to the plan you've defined?

reference/storage/management

Inevitably, data will be stored on multiple platforms throughout the data lifecycle. Consider the various tools used for data collection (mobile device, web form, paper, etc.), data storage (mobile device, laptop, centralized database, paper, etc.), data cleaning/analysis (excel spreadsheet, database, reporting platform, etc.) and ensure there is a safety and security plan for each of the various platforms used.

common ways of collecting data

There are many methods of collecting data, and each one comes with its own challenges. When choosing the data collection tools and processes you will use, think about how it can affect your message and what type of influence it might have when interacting with the participants.

Web-based surveys: There are a number of popular survey platforms available. From a functional perspective, it is important that they do what you need them to do, while not blowing your budget. However, you also need to ensure that they provide your participants with reasonably secure access (for example at least an https connection), so they are not transmitting sensitive information over unencrypted channels.

Moreover, if you are using an external provider, ensure that ideally they have no access to the data themselves, and cannot be legally compelled to hand it over without notifying you and allowing you to challenge that order. Refer to the data management chapter on more legal and practical considerations when working with external providers.

Mobile phone surveys: Mobile phone surveys are a great way to reach many participants fast with relatively low cost and a relatively low access barrier.

However, you are critically relying on existing infrastructure that is outside your control - the mobile phone network. That means, you are relying on telecommunications companies to follow responsible procedures, as all information passes through them; not to forget the huge amount of metadata that they will undoubtedly be in possession of: for example, the location the SMS was sent from, to whom, and times. They may be legally compelled to hand out either the metadata, or the data itself, to government authorities.

Consequently, you need to assess whether the phone network operator can be trusted, when thinking about using mobile phone surveys.

In-person surveys using mobiles or tablets: these surveys follow the same dynamics as interviews with regard to the need that it may be necessary to establish a safe space for the participant and situation of trust where they can give informed consent.

On the positive side, often you have more means of securring the data on devices, compared to working with paper. However, you need to ensure they are properly secured, in particular by encrypting their storage and establishing strong passwords. In addition to this, ensure there is a clear and secure way to process the data from the device into the management system, and check who else has access to the data stored on these devices.

quick checklist for tools, security and training

- Consider the pros and cons of using open source tools for the project (cost, customisability, independant audit, support)
- Do some features of the tools you have chosen present security challenges? If so, can they be managed or closed?
- Will data collectors need to edit their data/information after the fact? If so, ensure the technology supports the desired level of data editing, and ideally, that it records any changes made.
- Ensure selected tools allow for your desired range of data access permissions: will you need people with various levels of permissions to interact with specified sets of data? (see section: for your eyes only)
- Establish a response plan for potential emergency technical challenges (phones breaking, no wifi, etc.) and have support ready to go.
- Establish a response plan in case tools are confiscated or lost (eg. being able to wipe the data remotely)
- Consider using strong **encryption** software with good end-to-end implementation to protect data and communications content.
- Consider built-in protection features within your selected technology (e.g. device encryption) and their potential vulnerabilities
- Ensure staff have received appropriate training on the established informed consent process and are aware of ethical considerations around conducting research on other people
- Ensure staff have received appropriate training on data security for secure data management with devices that they will be using– password best practices, anonymised data, device security, etc.
- Have a process for reporting incidents, to provide care for your staff, as well as affected participants.

Note that this checklist is likely not to be exhaustive, but covers frequently observed blind spots.

working with existing data

Even if your project is not the one that actually collects and creates new data, there are a number of important issues to consider in terms of dealing with data responsibly. Once you have identified appropriate data sets that will support your project activities, there are **three primary areas of responsibility** to consider:

1. How the data is evaluated for use
2. How it is managed
3. How it is presented

evaluating existing data

understanding the context

While working with existing data, it is important to remember that data was collected by someone else for specific purposes and to address certain needs, which probably don't match yours. This can lead to **biases and gaps**: since the producers of this data already had an understanding of what they wanted and how they wanted it formatted and distributed, they may not have explicitly included information about the context in which it was produced. Unless you're lucky, nor will they highlight steps in the data cycle that were problematic. It is important to try and gain some insight and understanding of the context of the data before using it, through considering the following questions. You might not be able to answer all of them to the extent that you would like, but taking the time to work out which ones are essentials, and talking through the others with your team, is important.

- Is there enough **contextual information** documented by the originator of the data for you to be able to judge its usability? ie. do you know why it was collected, when, with whom, under which circumstances.
- Who **published**/licensed the data? Are they the same actor that collected the data? How much control did they have over the data collection process?
- Who **funded** the process? If the funder is known for pushing a particular agenda, bear this in mind when interpreting the data.
- What **biases** might the data hold? It may be useful to think of gender, geographic area, ethnicity, age, income level, internet access, at risk populations, etc

- What **methods** were used to collect the data? Where they appropriate? Are the limitations of those methods appropriately referenced in the way the data is presented?
- Are there any **social factors** that should be taken into account when interpreting the data?
- Is the data **inclusive?** Think of issues like gender, region, etc...
- Was **active engagement** ensured for the data subjects in the project design and data collection process? Was there an informed consent process? Was the data validated by data subjects?

managing existing data

Data should be usable as well as useful. Data will need to be machine readable for most types of analysis. Some datasets will not be in analysable formats, however, and would need to be transformed to be usable. This raises questions of **intellectual property** (am I allowed to modify the dataset?). More importantly, it introduces a transormative process (moving data from one format to another) that needs to be planned and executed carefully, so as not to lose information or inadvertently manipulate the data. Make sure to keep versions of the data, and clearly document your steps so that if you need to you can go back a step and see where you started from.

Some considerations for the licensing regimes and legal obligations governing existing data are:

- Will re-use or re-publication of the data violate any copyrights or other rights holders?
- How should these rights be considered in light of the responsible data challenges identified in the project risk assessment and in light of the project objectives?
- Do the licenses or permissions on the data limit the formats in which the data can be manipulated and reproduced?
- Does use of the data violate any terms of service?

who is hidden in the data?

All data reflects on people. Even when it is de-identified, anonymized or aggregated so that individuals are not immediately apparent, the possibility to re-identify individuals is always there, especially when combining that data with complementary sources of information. This is a phenomenon called the Mosaic effect: http://e-pluribusunum.com/2013/05/20/open-data-mosaic-effect/.

case study: “i know what you watched last summer”

> In 2006, the movie streaming company Netflix published 10 million movie rankings by 500,000 of its customers with the aim of developing a better way of ranking the movies that it recommends. The data was supposed to be anonymised by replacing customer names with random numbers and removing personally identifiable information.
>
> However, researchers cross referenced the Netflix Data with the public information available on the Internet Movie Database website – where users publically enter movie rankings under their own names. By comparing the two datasets for a small sample of users, it thus became possible to de-anonymise many of the people and understand their movie choices.
>
> Reference:
>
> “Why 'Anonymous' Data Sometimes Isn't”: http://archive.wired.com/politics/security/commentary/securitymatters/2007/12/securitymatters_1213

Even without identifying individuals, however, data about "things" or events can reflect the activities and habits of people and groups when combined with the right contextual knowlege. A dataset displaying water access points can, for example be used to infer information about population size, consumption habits and socio-economic status that could support discriminatory policies or exacerbate social tensions.

It's worth spending some time trying to identify types of people reflected by data, even when they are not readily apparent. Some areas to think about could include:

> - What does this data say about individuals? Who could be left out?
> - How could this data be combined with other datasets to provide more information?

> - Could this data be used to reveal any sensitive information about people or groups?

This exercise could be organized as a brainstorming excercise with the project teams, to try and work out as much as possible about the people behind a found data set, and then to use those intuitions to consider relevant responsibilities when using that data.

Collecting anonymised metadata on mobility of groups of persons on the basis of mobile phones may seem rather inocuous and certainly has helped making humanitarian aid more efficient in times of crises. However, depending on the context, this kind of information can have damaging effects on data subjects. As researcher Linnet Taylor puts it in her study of data science challenge involving African mobile data:
"(...) human mobility is becoming legible in new, more detailed ways (...) this carries with it the dual risk of rendering certain groups invisible and of misinterpreting what is visible. Thus this emerging ability to track movement in real time offers both the possibility of improved responses to conflict and forced migration, but also unprecedented power to surveil and control unwanted population movement."

Source: https://www.academia.edu/7502204/No_place_to_hide_The_ethics_and_analytics_of_tracking_mobility_using_mobile_phone_data

presenting existing data

When presenting existing data, it is critical to include information about the data's provenance and the context in which it was produced. This includes all of the information uncovered when evaluating and managing the data, but it also implies being transparent about what is **not** known about the data. Being explicit about the data's contextual unknowns is essential for enabling others to manage your project's data responsibly.

An easy way of thinking through the contextual information that might need to be presented is via the questions listed above, underneath 'Understanding the Context". You might want to present the answers to some of these questions together with your analysis of the data.

further resources:

Organizations that may be able to help you further on issues you may encounter:

- Geeks without Bounds - http://gwob.org
- Datakind - http://www.datakind.org

power to the people

Sometimes, it is easy to get caught up in the technical aspects of collecting and managing data, and forget that we are working with information about human beings and their lives. It is crucial to be responsible, open, transparent and inclusive in the data collection process, creating clear mechanisms for engagement and feedback, and always keeping in mind who is being described and who, ultimately, should have the last word.

representation

Data is a very powerful tool in telling a narrative about peoples lives and what their reality is. We need to understand how these narratives may impact people's lives, and be responsible in the ways we use and present this data. Given that it is important that data is used in the narrative responsibly, there are two aspects to consider:

Inclusion: data can never represent all people and all nuances of a cultural or social mosaic; no matter how big the sample size used, it is vital to remember that data is inherently exclusive and rarely, if ever, tells the whole story. If we ignore the very possible exclusions in data we risk perpetuating social inequalities. When collecting or finding data, ask yourself: who does the data leave out? Who is not reflected in this data and what impact might that have on their lives when using that data for programming, advocacy or policy change?

Accuracy - data needs to be as accurately reflective as possible of the reality of the people involved. Otherwise, the narrative that is developed from the data can be harmful for those people. The accuracy aspect is mainly about ensuring that the right questions are being asked: there are many elements of data accuracy that fall into the realm of research methods and techniques of data collection and so fall outside of the scope of this book. Bear in mind also that "accuracy" is inherently subjective.

consent

The concept of consent is so central, it occupies a full chapter in this book, and we will only mention it briefly here. One of the first steps involved in getting data is making sure that the people who are reflected in that data and who might be providing that data give permission for you to collect and use the data. This permission needs to be informed and given willingly, without feeling as though their other basic rights are at risk if they refuse consent. There are debates as to whether or not consent is always appropriate or achievable, and there are different levels of consent. Consent it a keystone to responsible data, and details on it are given in the relevant chapter.

ownership

Who owns the actual data, and what does that mean for the people involved? Many detailed laws and regulations exist, governed by legal frameworks around ownership of data. There is a growing debate on how traditional regulatory frameworks apply to the new data ecosystems. As these bigger debates settle, there are a few **ethical issues** we should think about that are broader than legal ownership. Ultimately, data belongs to the individual that the data describes, as they have the right to withhold consent and to remove thier data from the process.

However, there are a few problems with this: given the different types of data that exist, active data ownership is not always feasible. For instance, census data is often mandated by national law and an individual cannot opt out of being counted. Many forms of big data also have very complicated opt-out mechanisms, if they have any at all. This makes it very difficult for individuals to claim or retain effective ownership. Some basic questions to ask in this regard are:

- Does the individual have control over how their data is used?
- Do they have power to stop the data from being used in a certain way?
- Are they aware of how the data will be used and is there a way to communicate changes in data use?
- How are we accountable to the people reflected in the data? How is this accountability enforced?

There are also important issues to address regarding the power dynamics that are surfaced during the use of and collection of data. These are discussed in greater detail in the chapter on The world of data)

transparency

Transparency might seem to be the counterbalance to privacy and security; in reality the two concepts are perfectly compatible, and effectively two sides of the same coin, as long as they both work towards **empowerment and protection of people**. Transparency calls for disclosure: keeping an open and easily accessible account of methodology, intent of data use, who will use it and for how long, as well as how the data, the analyses and the learnings will be shared, and with whom. There should be a minimum list of data areas that can be safely disclosed. Specific instances in which this could prove to be important are donor reporting, monitoring and evaluating programs. Privacy should be granted to those without power: transparency should be demanded of those with power.

Information is power and as is commonly repeated, "with great power comes great responsibility". It is important to think of transparency in terms of how the data is being released, and whether all necessary steps to protect individuals have been taken into account. Is data anonymised and assessed for identification risks, or is the information for advocacy work where identities need to be disclosed? What is the most responsible way to navigate this scenario? Ultimately, **the responsiblity of making this choice lies with you.**

consent

Informed consent is the mechanism through which people agree to provide information for research or data collection projects. Generally, consent has been understood as something that is given by individuals during direct interaction with researchers or surveyors, and is composed of three components:

- **disclosure** of research objectives and any risks or negative consequences of participating
- **capacity** of individuals to understand the implications of participating
- **voluntariness** of their participation

Fully informed consent includes full disclosure of all potential risks and negative consequences of participation, and might not always be feasible or even possible. In some instances, it might be difficult for project staff to anticipate the way in which changing political or cultural contexts will influence risks for data subjects, and emergencies or pressing needs may in some cases justify data collection without the consent of data subjects. As a rule, however, projects which include data collection directly from data subjects should seek informed consent to the degree that is possible, especially in high risk contexts. Projects collecting data indirectly (through already existing data, or data scraping) will struggle to obtain informed consent of data subjects, but should consider these issues, and whether some form of mediated or simple consent is appropriate or possible.

Consent can be a highly technical and context-specific matter with many legal and regulatory implications. For example, consent in the context of academic research may be different from that in a healthcare context. Also take note that consent has a particular meaning in a legal context, which can vary across jurisdictions and across sectors (for a discussion of legal jursidictions and obligations, see the chapter in Managing Data).

This section is a high-level primer on how to think about consent when your project involves the collection, use and sharing of data. In that sense, it has a specific focus: it is not about consent to participate in the overall project to register for welfare benefits or receive medical attention, for example. It is about consent in relation to the use of the **data** that is provided as part of the overall development project.

A few words of caution:

- This section provides guidelines to help you think through the most common aspects of consent. It is **not a substitute for specific legal advice**. In order to comply with the legal regime applicable to a particular country, sector, type of data or category of person, you should talk to your legal or compliance team or specialist advisors.
- This section proceeds on the assumption that you have already thought through whether or not certain types of data should be collected at all, and have satisified yourself that you are able to safeguard what you do choose to collect. This section focuses on mechanisms to receive consent once you have already worked out whether, when and how to collect data.

elements of consent

Explaining further the essential elements of consent as mentioned above:

- **Notice or Disclosure**: the consent process has to ensure that the participant must be informed of the nature and purpose of the research, the procedures to be used, the expected benefits (if any), reasonably foreseeable risks, possibilities of not participating and procedures for confidentiality and anonymity.
- **Capacity, or Understanding**: the information must be easily understood in that specific context (language, technical jargon) of the participant and they must be given the opportunity to have any of their questions answered.
- **Voluntariness**: the consent to participate must be voluntary, free of any coercion or inflated promise, and not involve people who have power over the participants. In the development context, it is difficult - if not impossible - to remove any power dynamics entirely, however.

Additionally responsible data implies that we also include the following vital elements of consent:

- **Competence**: The participant must be competent to give consent and not incapacitated due to mental status, disease, or emergency.
- **Form of Consent**: In every consent process, the participant must signal agreement to participation, preferably in writing. In many cases written consent may not be appropriate and oral or implicit consent may do.
- **Waiver of rights**: Consent should not compel the participant to waive any legal rights or releases from liability for negligence.

what to disclose to participants/the data subject

Consent has traditionally been regulated as part of the ethical review mechanism for academic research, and there is no authoritative description of consent practices for civil society or development project data collection. Fully informed consent likely requires that projects disclose the following information to data subjects:

- **The nature and objectives of the project.**

This may include project history, partners and funders, as well as the specific objectives of the project and how those might change over time.

- **The purpose for which the data is being collected.**

How narrow or specific this needs to be might be dictated by the applicable legal regime, but in general, consent should be linked to a clearly identifiable purpose rather than a vague, "omnibus", open-ended one, however desirable the latter might be.

- **How the data will be used**

This includes how the data will be used internally and publically, and should also include how the data will and will not be shared, and any limits to use or the time for which data be stored.

- **What risks the data and its use might pose**.

This depends on the outcomes of the risk assessment and can never be exhaustive or authoritative. Disclosure will need to strike a balance between providing data subjects with the information they need to make a reasonably informed decsion about consent, and overloading data subjects with information that will have a negative effect on their engagement, and may itself present risks. For example, interviewees of the Belfast Project in the early 2000s were irresponsibly promised that the information they shared would not be accessible until after their death. They were not appropriately informed of the risk that this information could be accessed through legal channels. More information: http://www.timeshighereducation.co.uk/features/oral-history-where-next-after-the-belfast-project/2013679.article

- **What opportunities data subjects will have to review and influence the data** and its use over time, including opportunties to revoke consent.

how to disclose information to the participant/data subject

In order for people to be meaningfully informed, disclosure needs to be communicated in a way that is culturally, technically and socially appropriate. It also needs to be conscious of political and media landscapes and the fact that project contexts change over time. This has implications for the types of information that disclosure includes, as well as the mechanisms for communication disclosure and whether this takes place in-person, or through other types of media. Try to think about the following questions to help guide yourself through these issues:

- Are the methods of disclosure appropriate for the subjects in terms of language, media, and social contexts? Have you thought about accessible methods of disclosure (media/literacy/etc.)?
- Are there mechanisms built in to understand how project contexts change, and what implications that has for consent?
- Will you need to disclose information more than once or on an ongoing basis?
- Are people seeking the consent adequately trained to do so? For example, can the staff manage issues of consent and power within interpersonal relationships, and can they answer common questions that participants might have?
- When the processes of providing consent or disclosing information are mediated by technology, it is important to note the ways in which specific platforms and media influence the experiences of individuals. Has this been factored in?
- If applicable, how can the quality and consent received reflect on participatory projects? What does resistance to granting consent say about project relationships with communities and data participants? Do consent receipts suggest anything about degrees of engagement or the participatory nature of a project?
- If you are not including a consent process in your project, explain why? Does this put subjects at risk?
- Participants may change their mind about consent at various times, due to changing context or circumstances, and there always needs to be a way to individually revoke their consent. Have you provided for this, and have you made it easy and accessible so that the method of revocation does not in itself prove a barrier to exit?

consent in a digital world

Consent is fairly well understood in a traditional context. In a digital world that seems to offer infinite potential for information to be shared, adapted, manipulated and re-used, there is a need to re-callibrate our view on the 'traditional' concept of consent. It is necessary to include and think about principled and practical challenges for anyone using digital and mobile tools to collect, manage and share data for social impact. Equally, if data is collected offline, or if there is a combination of online and offline data, there may be issues around merging data, and human data entry errors that may get replicated when copied onto digital formats, as well as other technical or practical concerns.

One of the most challenging areas in which to understand the necessity and the limits of informed consent is in regard to crowdsourced or user-generated data.

consent for crowd sourcing:

> - Map potential risk areas for individuals/users in new and emerging crowd-sourced data technologies, information from which is used for advocacy
> - Identify, share and develop resources for mitigating these risks between organisations or initiatives that use crowdsourcing.
> - Develop best-practices guidance for organisations using crowd-sourcing technology for human rights research and campaigning.

developing a consent policy

To ensure that your project has an appropriate and realistic approach to consent, it is worth spending some time developing a general consent policy for your organisation or for a particular project. This will help with understanding what mechanisms of consent are appropriate, and to ensure that they are appropriately implemented. A consent policy should address the following components:

Risk Assessment: The level of consent and protection built in will depend on the risk assessment, including an assessment of urgency and data sensitivity. This may include some sort of threat modelling, or broader sense of assessing risks that may flow from wrongful or erroneous use or sharing of the data.

Disclosure: On the basis of the risk assessment, determine how much information you need to disclose to participants or data subjects. Generally this should include any potential risks or negative consequences that may follow from data collection and data use. It should also include detailed information about how data will be used, including

time and use limits, future disclosure and accountability, and sharing, as described below. Generally, there may be a tension between thorough disclosure and the willingness or ease for individuals to participate in data collection processes. This balance will need to be struck through a thoughtful conversation among project team members, ideally with the input of stakeholders and data subjects.

Time & Use Limits: Setting limits for uses and for how long consent is valid will need to be balanced by limits broad enough to allow for unanticipated activities and limited resources; these need to be checked against potential risk. In particular, you may need to build in triggers or system alerts for when data gathered for one purpose is likely to be used for another purpose, bearing in mind that you might need to go back to the original consent giver to renew or refresh the consent.

Transparency and Accountability: This is about how data will be used and how you will ensure accountability for that use. How will information on actual use be shared over time. Will information be shared on specific instances of data use or data sharing? How often will this be updated? Will this information be posted on media that is accessible and appropriate for people reflected in the data? Depending on jurisdiction it may be important to carry out a legal assessment of applicable data protection regimes, administrative regimes or other restrictive legislation.

Sharing: How do people who are using the data check whether the use is in-line with what they consented to? When gathering data, there is often an assumption that action will be taken upon that data; from the perspective of the participant, otherwise, why collect it? Other people may make use of this data, either with the organisation's consent or without, but it might not always be shared back with the participant. If people reflected in the data feel as though such uses are inappropriate or put them at risk, do they have mechanisms through which to raise these concerns?

implementing and maintaining a consent policy

The political and social contexts in which projects collect and maintain data rarely remain static. Regularly reviewing and implementing a consent policy over time is an important measure to ensure that it remains appropriate. It is also an excellent way to maintain engagement and input from data subjects, to forecast problems with data validity or support for data collection projects.

is consent broken?

A final word of caution: research increasingly finding indicates that consent has its limitations. For example, behavioural economics shows how easy it is to "nudge" people towards giving consent or choosing the riskier of options when faced when opt-ins or opt-outs.

Legal scholars such as Daniel Solove have talked of the limits of leaving users to regulate their own privacy and manage consents (in his words, "privacy self-management"). In his article, "Privacy Self-Management and the Consent Dilemma", he describes how consent is flawed for many reasons: due to bounded rationality, people cannot possibly understand what they are consenting to and what possible uses their data might be put to in the future; they are often not given an opportunity to withdraw once consent is obtained upfront; they cannot comprehend the multiplicity of actors with whom their data may be shared; they lack visibility of who those data handlers are, which means that they cannot approach them to request that their data is deleted or not shared further.

Further, the impact of aggregation - when small pieces of data about them are combined with each other to form a more complete picture - cannot be appreciated upfront at the point of consent, and the societal benefits that accrue from safeguarding privacy and anonymity should be not be waived through consent. In Solove's words,

> *"Privacy self-management takes refuge in consent. It attempts to be neutral about substance — whether certain forms of collecting, using, or disclosing personal data are good or bad — and instead focuses on whether people consent to various privacy practices. Consent legitimizes nearly any form of collection, use, or disclosure of personal data. Although privacy self-management is certainly a laudable and necessary component of any regulatory regime, I contend that it is being tasked with doing work beyond its capabilities. Privacy self- management does not provide people with meaningful control over their data."*

This does not suggest that consent is not meaningful. It is, and should be sought by projects collecting or generating data on people or groups, especially in contexts of power imbalance. More often than not, consent may be the most responsible way to implement a project with the knowledge and buy-in of the intended participants or

beneficiaries. Projects should be wary, however, of relying wholly on consent to legitimize any sort of practice or wide use of data without regard for the people behind it.

further resources

- A checklist for evaluating policies on consent is available at https://docs.google.com/a/theengineroom.org/document/d/1PJxBAP1rFkjq9p7NuYcN_G5iomfCML-qiMTn5SPPHxE/edit.
- Solove, Daniel J., Privacy Self-Management and the Consent Dilemma (November 4, 2012). 126 Harvard Law Review 1880 (2013); GWU Legal Studies Research Paper No. 2012-141; GWU Law School Public Law Research Paper No. 2012-141. Available at SSRN: http://ssrn.com/abstract=2171018
- Launching an SMS code of conduct for Crisis Mapping: http://irevolution.net/2013/02/25/launching-sms-code-of-conduct/

understanding data

verifying and cleaning data

managing bias and assumptions

verifying and cleaning data

In a perfect world your data would live in neat boxes, tagged, categorised and compiled, ready for perusal and analysis. Unfortunately, real life data collection is often messy, disorganised, uncategorised, jumbled and knotted. We might, for example, have online surveys with manual input, where the surveyors enter slightly different definitions of the same thing, making it impossible to clearly categorise, even if the two data points should clearly belong together.

Or, you might have a situation in which thousands of hand-written surveys are transcribed by dozens of volunteer college students, some of whom skipped a letter or a number here and there. The situation becomes even worse when the project involves collation and aggregation of already existing data sources, where we don't have complete knowledge of the methodology, or the data is provided in badly formatted files.

verifying data

Why is this important in the responsible data perspective? Ensuring the data you have collected is properly verified is vital. Without this the results cannot be relied upon for decision-making processes and you risk misrepresenting or doing harm to the populations you aim to help.

verifying data when you're close to the source

Integrity: Examine the raw data **before** doing anything to manipulate it. This usually includes making a complete backup of the original material, which is kept separate from any other activities (though it doesn't prove that the data has not been manipulated before reaching it's storage location). This can be vital at a later stage for ensuring accuracy, for availability (in case of accidental deletion) and for responding to accusations of manipulation, for example in a court situation. A number of methods are available to check that data has not been manipulated, depending on the method of data collection and storage. Working with paper, it is often better to work with photocopies or scanned images, while keeping the originals secure elsewhere.

Similarly, when working with digital data, you can:

> - Backup an exact copy. For example using, disk imaging software (http://en.wikipedia.org/wiki/Disk_image)
> - Verify you are using an exact copy of the original. For example using, file verification techniques (http://en.wikipedia.org/wiki/File_verification)

> For more on data integrity, please see the Data Management section on ‘A Home for Healthy Data’.

Go back and ask: Reach out to a small sample of the surveyed population for more in-depth analysis. This requires that in your collection efforts you have the ability to take and securely store data on the people you survey in order to contact them again (assuming it is safe to do so). If the initial group was part of a small qualitative research project, you might scale up through a more quantitative questionnaire to support your first research.

If you are close to the data collector, it may be possible for you to identify trends in the data and verify with the data collector. For example, if you see for a month there were 50 new data submissions, ask the data collector if that number feels right. Another

example, if you see some odd data patterns, you may be able to trace that data back to the original collector to ask them to explain how that data came about.

verifying data without the source

There are times where you are verifying data where you don't have access to the data collector/originator.

Cross-checking: If there are other data available about the same subject, look at how they compare. Do they reach the same outcome? Depending on the type of medium you have collected data with (video, photo, paper, etc), there are a number of techniques for remotely cross-checking the data. For example, with video footage, this can include:

> - Examining the extra metadata also captured in the data - such as time, location, camera type, length, resolution etc.
> - The content itself - such as:
>   - Is it possible to match location features such as roads, terrain etc. seen in the footage with other images like satellite photos, maps, regular photos?
>   - Is it possible to match identifying information such as building signage with other information on the area?
>   - Is the activity viewed or heard in the data validated by other information sources such other reports, social media, newspapers etc.
>
> - For further techniques, see The Verification Handbook http://verificationhandbook.com and the Citizen Evidence Lab http://citizenevidence.org/

cleaning data

Before moving further, we need to clean the data and prepare it for use by aggregating, filtering, reconciling and standardising our data and metadata.

Messy data may hide harmful information. If we don't make sure that we can clearly name, describe and recognise all the information contained in datasets, we might make improper assumptions about the risks that that data might pose. There is something to be said about having a birds eye view of an entire data set that is

organised and filtered, as opposed to trying to make sense of different points that are strewn in different directions. Gaps and assumptions come to the fore much more clearly after creating some order.

There are some very good sources available for learning what tools are out there, and how to use them: School of Data, for example, has a number of tutorials for data cleanup. From the School of Data - Data Cleaning course: cleaning data can mean a number of different things, including:

- finding and removing unwanted bits of data in spreadsheets
- formatting data correctly for the tools that you are using
- dealing with inconsistencies in the data
- structuring it so it can be used effectively for what you want it to do

HUMAN RESOURCES: The ability to efficiently manage large datasets is a science, and people qualified to do so are aptly (albeit not creatively) called *data scientists.* This person should have enough spreadsheet savvy to easily merge, filter and pivot complex data. It might be that your best solution is to hire a data scientist, or partner with an NGO that has in-house data science expertise. In any case, your project will benefit from you and your team gaining a firmer grasp on what procedures exist, and what can (or cannot) be done to clean up your dataset.

Filtering for the greater good

Any dataset can be described as a list of data points that are somehow related to one another. Usually these relations are described through two dimensions, because we visualise them on paper sheets or computer screens, both having only two axes: height and width. To make it simpler, imagine a spreadsheet: rows and columns are the two ways, or dimensions, that we have to relate and connect information. If a row represents one data point, like information about a high school teacher in Kenya, the columns will vertically categorize different types of information we have about all the teachers in the spreadsheet, like place of employment, salary, spoken languages, course taught, number of times the teacher has been absent in the last year, etc. Our goal with the data cleanup is to make sure all these categories show information that is relatable.

Filtering is usually the first step in checking data for consistency. Filtering creates a menu-like list of data aggregates present in a column, that let you choose to visualise only rows with specific content in that column. Tools like Microsoft Excel, Libre Office Calc, Apple Numbers or Google Sheets all offer strong filtering functionalities. A first filtering round will already offer an overview of how much the dataset is comparable and where the main need for cleanup lies. This is a hassle-free, accessible first step for you to understand more about your data on a very

general level, as having a clearer overview is useful to understanding how to move forward.

For more advanced data cleanup, **Open Refine** http://openrefine.org/ offers powerful ways to combine, compare and reconcile your data. Open Refine lets you combine and aggregate slightly differing types of data using algorithms that assist with "fuzzy" comparisons of not-quite-equal information. It also lets you execute **faceted search** on your data: a search combining multiple different filters and data points as facets.

For more in-depth information about data cleanup, visit "A gentle introduction to cleaning data" http://schoolofdata.org/handbook/courses/data-cleaning/, as well as "A gentle introduction to exploring and understanding your data" http://schoolofdata.org/handbook/courses/gentle-introduction-exploring-and-understanding-data/

preparation: describing your data

A complementary aspect of data cleanup is making sure the information that has been collected is fully and exhaustively described. Describing information means documenting and collecting information about information, also known as metadata (data about data). Typical metadata content types might be "date collected", "identifier", "size of picture", "format type" etc. When preparing the dataset for analysis, look into all the metadata you might have that describe your dataset and make sure they are documented.

While a great part of the description process will be directly connected to the data gathering phase, there are additional data points that can be defined during the preparation phase. Some might be automatic, like pulling metadata (timestamps, authors, file sizes etc) out of documents. Other might require manual work, like categorising by theme, type of response, etc - any information that might be helpful in the analysis phase.

formats and standards

Data can be described in many ways, and saved in a myriad of different file formats. It is smart to adopt common standards and file formats for the data, so that they can be more easily shareable, more resilient (future-proof) and also comparable with other datasets; or, interoperable. One example of a standard file format is the **comma-separated value** format (or .csv) for spreadsheets. While each spreadsheet software has its own proprietary file format that usually also provides additional software-specific functionalities, the .csv has the strength of being shareable across platforms, and a strong open description of its formatting so that the data lives in a format that can be adapted to other purposes. Having the data live in a proprietary file format like .xls might mean that one day the company that owns it stops supporting it, the dataset becomes unusable, and the knowledge is lost.

managing bias and assumptions

can i trust what my dataset is telling me?

At this phase, you have already collected your data, cleaned it up, described it and standardised its formats and inputs (if you haven't, you might want to have a look at the previous chapter on verifying, cleaning, and preparing your data. Your dataset is begging to be analysed. It might be chock full of interesting informational clues that you can't wait to pull out and present. You have a distinct feeling that the dataset you have is going to answer your questions: more importantly, you are pretty sure it will give you the answers you want.

But wait! Before analyzing your data, this is the right moment to examine **your assumptions** (and those of others) about the dataset. There are numerous pitfalls when trying to answer questions from data. In this chapter we will discuss various challenges that may arise when looking into the collected data. The optimal situation would be that these considerations have already been made during your design phase, but in real life you are often faced with data already collected by someone else, leaving you to make sense of it. We aim to cover both situations here.

This chapter explores the following questions:

- How do I make sure my data is accurate?
- How can I make sure I understand what "accurate" means?
- What are responsible ways to remove noise from my data?
- What is causation, and when can I talk about it?

making sure your data isn't biased

All data, no matter how it is collected, will contain a certain amount of bias. It is your task to analyse what those biases might be, to minimise them to the extent possible, identify the ones that cannot be removed, and make sure that persistent biases are well known and explicitly flagged throughout the research. Throughout the data cycle you should keep a keen eye on eliminating as much bias as possible.

Bias is the tendency of results to 'favor' a certain outcome, due to the implicit construction or logic of the collection or processing of the data, the way that the data was collected (setting, sequence of questions) and the way that the data is analysed.

Below we will address some specific points that are useful for spotting red flags in your data.

Some data is collected using a **sample** of the universe or total population, on the basis of which we make generalisations on the phenomenon. Being aware of sampling basis, how it impacts analysis and limits what we can say about the overall population. Was your sample sufficient to allow for certain types of conclusions? Did it bias your analysis and findings in any way?

In the planning and design phase, bias can be introduced when focus issues and topics are selected. In the data collection process we can have **response** bias due to the phrasing or sequence of the questions asked or the setting of the data collection site. One issue is **cultural bias** where some answers are more socially desirable then others, and may skew the result.

You should take special care if you are working with **comparative** data: differences in collection techniques, or even differing definitions, might seriously skew your results. If, for instance, you are collecting and comparing data on sexual abuse between several countries (or even provinces) terms like assault and rape might mean very different things. These differences grow larger as language and culture differ.

Consider testing for **data collection** bias in your results. Often results may be accidentally or deliberately manipulated by the people collecting the data. For example, participants may fear losing their benefits if they give negative answers to the

person asking the question. Similarly, the collector may have had trouble in gaining access to the correct mix of participants. It is possible to check for this by taking a small sample of the results and re-validating the data (see the previous section).

Every dataset contains **outliers;** datapoints that are so different from all the others that it really skews the results. These can be anomalies (one rich person living in a village) or may just be errors in data entry (entering a few extra "0s" to someone's income). Going through the dataset carefully and removing these is part the data analyst's standard toolkit. However, it is important to make sure you are only reducing data noise, not changing the data to fit your expected outcome.

Correlation vs causation. Even if two variables might seem to be related, it doesn't mean that one caused the other. The classic example used here is the correlation between the rise of crime rates and ice cream consumption during summer months in the US. The two variables are correlated, but nether causes the other - both are, in fact, linked causally to temperature, but not to each other.

Why throwing away some of the results might actually improve the accuracy of your data

If your data is derived from a sample population, you might have inadvertently picked one or two individuals that are way off the charts with some of the parameters, in such a way that you can not generalise the results. For example, in measuring the results of an income generation project, you have three women who have incomes 3 times more then all 200 others: including these would significantly change the results of your data and may have been caused by a simple data entry error.

If the data has been manually entered or been automatically collected, the outliers might also derive from measurement errors, or a typo.

Removing such data actually makes the remaining data more meaningful (and less noisy), and provides a more concrete and realistic dataset.

Compare with other data and analysis: Are there similar, comparable data collection efforts from other countries or groups? There are a number of resources online which can be useful data sources. For example, the IATI Registry (http://www.iatiregistry.org).

Go back and ask: reach out to a small sample of the surveyed population for a more in-depth analysis. This requires that in your collection efforts you have the ability to take and securely store data on the people you survey in order to be able to contact

them again in future (as long as it is safe to do so). If the initial group was part of a small qualitative research project, maybe you can scale up through a more quantitative questionnaire to support your first research.

Connect with experts in the field for opinion: Experts working in the field you are doing research on might provide a valuable resource in knowing what your results actually mean, and whether there are gaps or blind spots in your research that you should address before starting the analysis.

sharing data

when to share, when to publish

sharing data

publishing data

anonymising data

presenting data

when to share, when to publish

Once data is collected, it can be released in a variety of ways, from closed networks within an organisation, to platform dependent sharing between peer organisations, to publishing with closed licenses, to publishing with fully open licenses. It's often tempting to think that making data available to the widest possible audience is the best way to maximize that data's impact. This approach can also be motivated by the desire to do justice to all the hard work put into collecting, cleaning, verifying and analysing. However, it's worth carefully considering the various forms that sharing can take.

Models for sharing can differ both qualitatively and quantitatively, and there are various levels of sharing that you can adopt. Sharing may be an optional or a mandatory activity, depending on the source of funding or sponsorship, the nature of the organization, the type of data involved and other factors. It can include information about the original exercise and purpose for which the data was collected, but can always be repurposed by others, sometimes for purposes contrary to the project that first shares.

benefits and limitations of sharing

There are many obvious benefits to sharing data. Doing so can maximize the impact of data (or conclusions drawn from it), inform collaboration, provide stronger evidence for advocacy, increase efficiency of service delivery among a wider audience than just within your project, or play a role in decision making within other projects, to name just a few potential benefits.

Publishing your data can also allow people who might not have otherwise been well-informed enough about your project, to have a say - for example, those who are reflected in the data. Without the data being made 'open' and accessible to them, their information channels about what is happening in their communities might be severely limited: put otherwise, providing them with proactive access to information is a crucial step towards empowering them to make their opinions known.

Sharing, in the sense of publishing open data, is also an increasing trend. The open data and open government movements, as part of pushes for better transparency and accountability, as well as recent interest in shared measurement for project evaluation, are just some examples of how the international norm for sharing data has gained powerful traction in recent years.

Sharing can also have unintended consequences, however. Once data is published, it's impossible to anticipate how it might be shared further, and once it's out in the open, there's no telling how it will be adopted, re-purposed and re-rused for any number of purposes. These might be positive purposes, finding uses for your data that you had never imagined yourself. But, some of these purposes might be malicious or run counter to your project's strategic objectives, and others might call into question the premises on which your project work was conducted, still others might expose your data to new audiences and new risks.

Given our inability to see into the future, it's especially important to think carefully about what kind of data gets shared, and the relationships, licenses and agreements that govern limited sharing. Apart from the ethical implications of unforeseen use by others, there may be practical considerations: for example, participants seeing that their data is used in a way they don't agree with, or that puts them in danger, might mean that they refuse to participate in subsequent research or development efforts,

either with you or with other data-intensive projects generally. As you can imagine, this has much wide societal consequences, and deserves careful thought and work to avoid such a situation.

Sometimes, technical measures to strip identifiers, redact sensitive information or otherwise "anonymise" data may be sufficient to mitigate against such potential harm. De-identification is problematic, however, and rarely works as a magic bullet. For a thorough discussion on this, see the section on anonymizing data.

The bottom line here is that you should carefully consider the implications of sharing (from the point of view of the people to whom the data relates, as well as the bigger picture), whether to share at all, and the licensing conditions or terms and tools that you can use to reduce the risk of harm, while still permitting beneficial outcomes.

whose data is it anyway?

Many will argue that data should belong to the people who have provided it through reporting, answering surveys or simply by using devices and media that generate a data trail. This implies that *data subjects* have a right to be informed and consulted about how their data is used, and to require that their explicit consent be obtained for specified purposes and be "refreshed" for each new purpose beyond the scope of the original consent. However, this norm is difficult to operationalise in many situations: for example, among communities with low levels of data literacy, or in particularly rural and hard to reach areas.

Others argue that data subjects have rights associated with the **result** of analyses conducted on the data, in addition to with the raw data itself. These sorts of "secondary" data rights may suggest entirely different ways of engaging with data subjects and participants, and the ways in which consent is operationalised.

It will inevitably be up to individual projects to determine what kinds of data rights are appropriate in specific instances. It will be important that these decisions are made explicit when data is shared or published, however, and that appropriate licenses or agreements are applied. (See chapter: Power to the People)

legal and contractual frameworks

Legal systems often include some version of a "purpose limitation" principle, in which data collected for one purpose cannot be used for any other purpose without the consent of the data subject. This is often seen as a way of respecting boundaries and choices.

Data sharing can often challenge this foundational principle. If, at the data collection stage, another common principle of "data minimisation" (collect no more than you need for that specified purpose) was also ignored, this problem is compounded. Having collected too much data is problematic, but potentially unproblematic, as long as the data is kept in-house. This is another reason you need to think carefully before you publish or share more widely however- see section Sharing Data.

For sharing data with a limited set of actors it is worth considering whether data sharing should be governed by explicit agreements such as MoUs or even contracts. Agreements are imperfect solutions insofar as enforcing them is rarely simple, and breach means that the data is already "out of the bag", but they also have some advantages. Entering into agreements about the conditions, limitations and ethical guidelines that govern data sharing can impose some measure of control, and can also establish a shared set of expectations and surface previously unforeseeen risks. Highly explicit agreements about risk and responsibility can also be shared further down the chain of actors with whom data might be shared and reinforce awareness about a responsible data approach among actors not directly within your project's sphere of influence.

sharing data

After considering the points in the previous chapter, you may have come to the conclusion that your data can be responsibly released into the wild, at some level. The next sections are designed to help you think about the different groups you can share with, and the appropriate checks and balances you should consider for each level of dissemination.

sharing internally (within your organisation)

Perhaps the data that you have collected will be of particular interest to your colleagues, who might be working on similar issues, or within the same country or region. This is a relatively restricted level of sharing - just within your own organisation, department or formally organised consortia or partners.

Be aware of the blind spots that internal sharing often entails: you are probably asking like-minded individuals who know where you're coming from, understand the issues you're dealing with, and know what to look for. They will be able to provide you with vital feedback, validation of assumptions, and will understand the importance of keeping the data safe and secure in this very sensitive first phase. But they will likely be sharing your own biases and won't be able provide truly independent feedback.

The wider you go within your organisation (or consortium/partnership), the more diverse feedback will be, bringing with it truly fresh perspectives come in. However, the wider you go, the less control you have. An important aspect of this level of sharing is the presence of an **organisational security policy**: a set of rules and guidelines providing you with a framework that you can safely assume your colleagues, partners or co-workers, will follow when dealing with the data.

SEE: data consent, for your eyes only...?

TOOLS: collaborative tools, private Github repos, etc

controlled closed sharing

Another sharing option might be to share it with people outside of your own organisation, but to retain some level of control. It's worth bearing in mind that for most NGOs, this almost always has to be a smaller and tighter data set than what can be shared in the earlier stage: this will be easier if you have already had input from those fresh perspectives we mentioned above.

For example, the data could be shared with peers for external opinion, or even data aggregation, in case they are working on similar themes (beware of combining apples and pears when combining different datasets, however).

Data aggregation is a crucial point of risk assessment: sometimes standalone data deemed safe becomes harmful when combined with other datasets, or data that you thought was anonymised becomes easily discernible once combined with other data, using triangulation techniques. (see section: anonymising data)

Another reason for sharing externally is the advantage of getting different expertise. Sharing data with other organisations can help to recognise the gaps in your dataset, making for a more resilient and more trustworthy dataset. It is all too easy to unknowingly apply your own personal biases when collecting data (for example, within questions asked in surveys, or structuring of the data), and sharing it with people who weren't involved in early stages might help to identify these biases.

One last aspect to mention here is the **sanity check**: sharing your data with unusual suspects will provide you with opinions that are outside of the echo chamber you are used to and help expand the dimensions of your project. You could, for example, share the data back with the people reflected in the data, and see what they think of how you've structured it. Often, doing this is a responsible thing to do anyway.

When connecting with external organisations, it might be useful (or sometimes, organisationally mandated) to put in place legal constraints of how your data will be used. Tools like a Non-disclosure agreement can provide legal leverage for ensuring your data is not misused, or that it is treated safely, and doesn't get shared any wider than you are envisioning.

When collaborating using the same tools, software will often allow you to configure access permissions or check logs for suspicious activity. This will not help you prevent leakage, but will help you identify them proactively, which will in turn allow you to take

measures for containing damage.

TOOLS: Non-disclosure agreements, Memorandum of Understanding, collaborative software

the point of no return

So you want to **publish** your data - at this point, you're most likely talking about a smaller, more controlled portion of your data, which you have carefully checked for any weak points, sensitive points, inaccuracies, and biases.

Once the dataset is shared publicly, the proverbial cat is out of the bag. Any weak point in the information, any personally identifiable information that hasn't been properly addressed, will be impossible to mitigate because someone might have already made a copy. The sections above offer a strong validation and checkup process; however, a **final risk assessment** of the data is more than warranted.

If you're at this point, your dataset has likely changed since the very first version you might have been working with, and is hopefully more robust and secure now, it might also be that some aspects have slipped through the cracks.

There's another group that need to get back to at this point: the people who are reflected in the data, or **the data subjects**. Before publishing information that, however unlikely, might put individuals at risk of harm, you should set up procedures to connect either directly with the involved people, or with representatives of communities you are collecting data on, to get their final go-ahead before data is published.

If you are confident that the data is ready for sharing with the world, please proceed to the next chapter: **publishing data.**

TOOLS: Risk mapping tool in development, from the engine room

publishing data

So you want to publish your data - here we'll take a look at various options for formats, platforms to use, licensing and whether or not to make the data 'open'.

open vs closed

You might have heard of the buzz around 'open data' - but what does this actually mean?

> According to the Open Definition (http://opendefinition.org/):
> Open data can be **freely used, modified, and shared** by **anyone** for **any purpose**

This means that **published** data, or **online** data is not necessarily **open data.** For example - data that is published as an Excel table within a PDF document, without an open license (more on this below) - is not open data, because it can't be easily managed or re-used. Whether or not you want to 'open' your data is an important consideration.

There are many benefits to open data:

> - It can be easily re-used and re-purposed for complimentary development and social good activities, saving resources and avoiding duplication.
> - It can be a means of building capacities and standards for evidence among development and non-profit organizations.
> - It encourages transparency and encourages accountability to participants, beneficiaries, peers and data subjects.
> - It can be verifed and quality controlled by a larger group of interested parties.
> - It can also be easily combined with other datasets to address longstanding problems, and notice patterns that might not have been obvious in isolated data sets.

However, making data 'open' also allows potentially malicious actors to use the data for their purposes: this is really important to bear in mind. As the saying goes - the best thing to do with your data will be thought of by someone else... but this also means that, potentially, the worst thing to do with your data will also be thought of by someone else.

case study: school attendance

In Country X, there was a region where, in certain schools, the school attendance of girls aged between 12-15 was unexpectedly low. The Ministry of Education ran advocacy and information campaigns to families, trying to highlight the importance of girls' education - but there was no change. The problem was only solved when open data sets from the Ministry of Education and the Ministry of Health were combined: it turned out that all of the schools that didn't have proper sanitary facilities were experiencing the drop in girls attendance. When the girls started menstruation, they had to stay at home, and this led to a big drop off in their attendance after they had missed so much school so regularly. The Ministry of Health installed proper sanitary facilities in the schools, and the girls attendance returned to normal levels.

licensing

Regardless of whether you choose to make your data 'open' or not, there are a number of different licenses that you can use for your data. Some aren't included within the Open Definition, because they specify what kinds of activities (eg. for commercial purposes) the data can be used for, and according to the Open Definition anyone should be able to do whatever they like with the data.

It's important to think about licensing for a number of reasons: firstly, **if you don't license your data properly, others won't know if or how they can use it in their work**, so they might either have to get in contact with you to check first, or they might simply decide not to use it. If you've decided to put the data online to help others in their work, this would be a shame!

Secondly, **licensing makes it clear how you want your data to be used.** You might want to retain rights to attribution on it, or simply let people do whatever they want with it. If you choose to retain attribution, it makes it easier for you to see how your data has been used by other parties - this can be interesting to demonstrate the impact of the data that you collected (for example, has it been used to support other development efforts?) - or to identify uses of it that you might not have thought of. Keeping tabs on your data also means you are maintaining some level of accountability to the data owners.

Types of licenses
Broadly speaking, there are two main fields of license you can choose from: Creative Commons Licenses, and the Open Data Commons licenses (http://opendatacommons.org/licenses/odbl/).
Creative Commons Licenses are more commonly used for content rather than raw data - for example, photos, text, reports. They have a nice online 'License Chooser' which will help you pick the right one for
you: https://creativecommons.org/choose/.
The Open Data Commons licenses were specifically created for databases, so might be more relevant here. There are two basic options: Public Domain Dedication License, or PDDL, which puts all of your material in the public domain (ie. for anyone to use), or the Share-Alike (plus Attribution) option, the Open Database License (ODbL). Here's a set of Frequently Asked Questions about the Open Data Commons licenses (http://opendatacommons.org/faq/licenses/#General) and here is some more information on Open Data Licensing
http://opendefinition.org/guide/data/.

publishing to the iati standard

The International Aid Transparency Initiative (IATI) represents a tremendous push for open and transparent data in the international development sector. Initially created as a process through which donors and partner governments could publish their aid flow data, it has quickly become a data standard through which a number of organisations (donors, large and small NGOs, recipient countries and governments) can put their data online in an interoperable way. (More about IATI
here: http://www.aidtransparency.net/about.)

There are a number of reasons that organisations should be enthusiastic about publishing to IATI. It's an important moment for transparency and accountability, and an opportunity to strengthen organisation's data practices. There are risks to consider as well, however.

Recently, there has been a push for 'traceability' and increased granularity within geocoded data - ie, more detail on where exactly projects are taking place. This would, in theory, allow for more people to keep a closer eye on where exactly money is being spent - is it going where it is saying it is going? But pushing for more granular geocoded data might, in some cases, require a risk assessment first. For example, if an NGO in a country is doing projects that are seen as 'troublemaking' by the government in power at the time, providing them (or anyone) with details on where exactly the projects are taking place might put project implementers in danger.

case study

> In an effort to increase transparency and reduce corruption, the government of a small African country are thinking about making it mandatory that **all** NGOs working in the country publish to IATI. This would include project documents, details of financial budgeting and spending, and potentially geocoded data of where project activities are taking place, too. By employing a 'publish by default' strategy, they want to show that they are truly committed to rooting out financial corruption in the country to international donors. However, the government also has some restrictive human rights policies in place - and also has a history of political instability, meaning that those in positions of power change frequently. Activities that are deemed 'suitable' by NGOs within this political climate could change suddenly, leaving those NGOs and their constituents in danger by publishing their activities to IATI. They use the argument that organisations can ask for exemptions if they don't feel comfortable publishing their activities: but in many of these cases, asking for an exemption **not** to publish might be interpreted as a signal to government that they are engaging in "unlawful" activities. So here, 'publish by default' has some hidden, potentially dangerous consequences.

While increasing transparency and encouraging accountability around aid activities is clearly a good move, it is crucial to bear in mind the potential risks and outlying cases through which publishing to IATI might put people in danger. The IATI community is made up of people who have vast amounts of knowledge about the data produced and used by global development organisations across the world - if you have a person tasked with publishing your organisation's data to IATI in your organisation, it might be worth sitting down with them to discuss exactly what is being published, and what is not.

TOOLS: Document Cloud, AidStream, Github, Google Docs, CKAN, HURIDOCS

anonymising data

Data can have real consequences for real people, and often these consequences are as unintended as they are harmful. This is regularly the case when Personally Identifiable Information (PII) is published, or when seemingly innocuous data is mashed and collated with other datasets. This is why it is very important to try and anonymise information before publishing in any way at all. However, there are many cases where efforts to anonymise have also failed.

At the end of the day, there might be no such thing as perfect anonymization. Removing all personally identifiable information from a data set isn't necessarily enough to protect identities in that data set. With the increasing sophistication of analytical techniques and algorithms, "de-identified" data sets can be combined with other supposedly anonymous data to re-identify individuals and the data associated with them. This phenomenon, known as the Mosaic Effect, is particularly challenging because evaluating the risk of it occuring requires one to anticipate all the different types of data that exist or may be produced, and which could be combined with that data set, which simply isn't possible.

As such, this chapter will talk about strategies for de-identification, because that word doesn't have the "magic bullet" feeling that is often associated with anonymization, and because the word itself implies that it can be undone.

It's also important to underscore that de-identification is a complicated and imperfect science. This chapter aims to help you navigate different techniques, but be careful: Don't take decisions on publishing lightly, seek expert advice and err on the side of caution.

Once the data is published, there is no turning back.

what data should be de-identified?

The short answer is, any and all data that has any potential to identify a person But it is still important to conduct a thorough risk assessment of the possible consequences of data release during project design and update it before release. Once the data is out, it's too late. Consider the following:

- Can an individual be identified from the data, or, from the data and other relatively accessible information?
- Does the data ‘relate to’ the identifiable individual, whether in personal or family life, business or profession?
- Is the data ‘obviously about’ a particular individual?
- Is the data ‘linked to’ an individual so that it provides particular information about them?
- Is the data used, or is it to be used, to inform or influence actions or decisions affecting an identifiable individual?
- Does the data have any biographical significance in relation to the individual?
- Does the data focus or concentrate on the individual as its central theme rather than on some object, transaction or event?
- Does the data impact or have the potential to impact on an individual, whether in a personal, family, business or professional capacity?

If the answer is "yes" or even "maybe" to any of these, you need to anonymize!

common types of personally identifiable data

Information that at first glance seems not-identifiable may become so, especially when combined with other data and subjected to powerful algorythmic analysis. This list gives examples, but is nowhere near comprehensive. Do evaluate for your context.

- Age
- Ethnicity
- Gender
- District/County

- Highest Level of Education
- Medications prescribed
- Geo-location of login

thinking about risk

Once you have done your data sensitivity assessment, you need to do a risk assessment of the context. This gives you an idea of how dangerous the data could be for the data subjects if it got out, and how much danger there is for that. Some of the things you should think about are:

- What might be the consequences be for individuals or groups in the de-identified data if it is reversed?
- What people or groups may have an interest in trying to re-identify your data? (Intelligence agencies, hackers, curious data scientists etc.)
- What other datasets are available that may result in re-identify the data you are releasing?
- What is your release strategy for your data? (For example, how is it being released to media? Is it possible that they may accidentally/deliberately add identifiable data?)
- What technical and version control methods are you going to use? (For example, to ensure you release the correct anonymised version of your data)

anonymisation techniques

These are some techniques that help in anonymising data, once the process above has yielded that it may be advisable to to do so. For more detailed descriptions of these techniques, their limitations and in which contexts they are most appropriate, see the Anonymization guide produced by the UK Information Commissioner's Office (http://ico.org.uk/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf).

- Data masking: This describes supplying only part of a dataset (e.g.. taking out columns from a spreadsheet) or deleting these parts from the dataset completely.
- Pseudonimisation: This describes the exchange of values for codes - this way, for example a name might be replaced with a number, but the same number will show up in every instance where the name was.

- Aggregation: Instead of providing the raw data, this would aggregate especially small amounts of information; rounding large numbers, or providing only small samples of larger datasets.
- Derived data: This describes a process where the original intent is kept, but the output changed. An example would be to provide the age of a person instead of the exact date of birth.

case study: new york taxis

In 2014, New York City released data under a Freedom of Information Request on 192 million taxi trips and fares made the year before. It held data on items such as pickup and drop off points. This potentially had a lot of useful research benefits, such as city planning. It also contained personally identifiable information about the name of the driver, taxi license and taxi plate number - this information was supposed to be anonymised. However, this anonymisation was very soon worked around, leading to a lot of intimate information being available online, about both drivers and passengers.

While an effort had been made to hide these pieces of information using a method known as “hashing,” it was undermined by a poor understanding of how it works. For example, due to the fact that taxi license numbers are assigned using a specific six or seven digit method, the anonymising method was weakened because it was limited to only three million possible combinations. It then took only minutes using a modern computer to reverse the anonymising method and reveal taxi license numbers. Due to the fact that the NYC Taxi and Limousine commission also provides data linking real names to taxi license numbers, researchers could get to the name of the driver.

The result was that it was possible to figure out who was the driver of nearly every one of the 192 million journeys. From this it was possible to determine how much money each driver made, where they lived, the dates and times when they were working and in what areas they worked. In some cases, it was also possible to infer journeys made by members of the public, in particular celebrities, or visitors of stripclubs.

Sources:

On Taxis and Rainbows: Lessons from NYC’s improperly anonymized taxi logs

https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1

> NYC Taxi Data Blunder Reveals Which Celebs Don't Tip—And Who Frequents Strip Clubs
>
> http://www.fastcompany.com/3036573/fast-feed/nyc-taxi-data-blunder-reveals-which-celebs-dont-tip-and-who-frequents-strip-clubs

get help

It can be helpful during this process to create an expert group of advisors with experience on both your project and with releasing data in a safe manner. Ideally some of these people will be outsiders, with a good knowledge of the areas where you work but without potentially biased connections to the project. The overall objective of such a pre-release review is to review all possible negative scenarios that might occur because of the data.

Judging by many previous failures, it makes sense to also seek advice from people who have skills relevant to de-anonymising and de-identification, such as data scientists, open data experts and ethical hackers. Pre-release reviews should also include increasing your protection methods in light of the potential increase in attention you may be subject to afterwards. For example, if original datasets are still not secured properly (physically or digitally), now is the time to ensure you do this. After release, you may find it becomes harder to do this. Especially if adversaries are technically adept.

further resources

The UK Anonymisation Network provides consultation and training for NGOs.

Anonymization guide produced by the UK Information Commissioner's Office (http://ico.org.uk/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf).

presenting data

power and the visual

There are many ways of presenting data: from infographics to narrative reports, case studies and long form investigative articles, to graffiti or conceptual art. The list goes on, but what's important is that different mediums bring different ethical and moral challenges with them when transforming data into message.

No data is neutral, and presenting information can surface even more biases. Use data as accurately as possible, and take care not to misrepresent topics or skew the data in order to suit your particular advocacy or campaigning need - it will make for an unreliable and easily discreditable campaign, and it will probably put your organisation's reputation at risk. More importantly, you are breaking the agreement with the people you are representing.

Visual and narrative manipulation and misrepresentation of data, be it by chance or on purpose, is an issue of great importance. Publishing your data is often the last mile, and can be the culmination of the whole project. Digging deep into visual wizardry is outside of the scope of this book; however we warmly suggest reading up on the subject before going public. A very good starting point is Mushon Zer Aviv's short essay "How to lie with data visualisation": https://visualisingadvocacy.org/blog/disinformation-visualization-how-lie-datavis

power and representation

In most initiatives dealing with development, vulnerability and marginalisation are critical issues for the people we work with. As with other elements concerning data, inclusion of marginalised people is something we have to be sensitive to. Examples of these groups could include women, persons living with disabilities, LGBTI communities, people from certain geographical areas or people of different racial, ethnic and cultural backgrounds.

When presenting data, ensure that these groups are appropriately treated and they are not being omitted, trivialised, judged or romanticised. Be intentional in how they are represented and what this means for your participants and beyond.

case study: "i know where your school is"

A local NGO was involved in investigating and documenting human rights activities. They used the results to pressure their own government. One of their initiatives was to interview people on camera and produce a documentary about the country - seeking to give voice to victims. To protect the participants, they took steps to keep their identity secret and told them their story would not be traceable to them as individuals. The NGO subsequently put this documentary online to expose the issue and feed into their other advocacy activities. However, they had little experience in editing video and and when they release the footage, faces were blurred but not sufficiently (e.g. when people got up or moved on their seats, their faces would be visible for a few instances) and it had not been considered that victims speaking about what exactly perpetrators had done to them would mean they could be identified even without their name having been stated (since at a minimum, the actual perpetrator could identify their victims when testifying a certain level of detail). They also failed to consider the danger in the footage of easily recognizable location features (such a school signs in the background, or buildings which can be easily found through Google Maps or Earth and lead to identification). In at least one case, an interviewee who had given testimony was re-victimized by subsequently getting detained by government forces and badly beaten.

Lessons:

- Do not assume that the consent of the person you are documenting also means you do not need to check if they fully understand the risks and potential implications
- Your staff should make the final decision if it is safe to release or not and they should have been trained and empowered to do so responsibly

(Note on not patronizing people: it is important to respect people's agency while minimizing risks to them. So it may not be your call to make to prevent someone from telling their story - even if it entails risks - as long as the person fully understands that. It *is,* however, your responsibility to make sure that when they think they will not be recognizable, they really are not!)

> Mitigation:
> - Consider potential risk of the use of information you collect
> - Make sure to consult the people who actually understand the context and what sensitivity means in it

further resources:

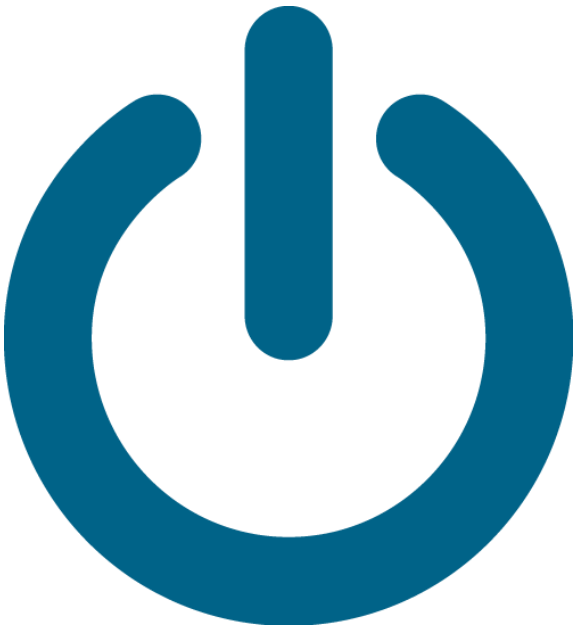
TacticalTech's Visualising Information for Advocacy guide: http://visualisingadvocacy.org

School of Data's Visualizing Data course: http://schoolofdata.org

Mushon Zer Aviv's short essay "How to lie with data visualisation": https://visualisingadvocacy.org/blog/disinformation-visualization-how-lie-datavis

closing a project

project closure - what happens to the data?



project closure - what happens to the data?

Projects end for many reasons. Funding cycles may close. Project goals might be achieved. Even if a project has not ended, the data lifecycle may end during a project, when data no longer needs to be collected and/or referenced in the project.

In all instances, projects that have a data component must address the end of the data lifecycle. **Deciding what to do with the data after it has served its purpose is an integral part of the project design process and should be informed by the risk assessment**. Usually, this will encompass the following:

- Mapping where the data is and ensuring it is only in places where it should be
- When collected data has been be shared (internally or externally), ensure that responsible data procedures are upheld after closing
- Disposing with (parts of) the data or archiving it, either internally or externally

Where possible, decisions about data sharing, archiving or disposing should be informed or driven by participants depending on project type.

what data do i have and where is it?

Think thoroughly about what types of data live where, in order to plan for proper project closure.

Be sure to consider all the places where data might be collected/stored. Data can live on paper, mobile devices, centralised (cloud-based) databases, distributed databases (Excel/Access/other applications across individual laptops), etc. Think about whether some of your colleagues might have copies of the data or parts of it on a private device, for example a mobile phone. Remember also that anything you have ever emailed may still be on the mail server - who does this belong to?

You may have gone through a data cleaning process to responsibly share and publish your data, but raw data and metadata may still exist on personal laptops or mobile devices.

tips on hosted data

> If you are using third-party applications for data collection or storage, your data is likely hosted by an external provider. It will be important to work with the external provider (**before** you sign a Service Level Agreement) to plan how data will be archived or disposed at the end of the data cycle.

If you are using open-source applications, you probably also host the data within your organisation (but you may also be using an external provider). Either way, ensure you know where your data is and that you have a process agreed upon with regards to how to dispose or archive it.

Data archiving and disposing is not always a simple process. It will be important to plan the time, resources and budget necessary to responsibly protect your data through the end of the lifecycle.

what if i want to keep the data forever?

A good principle developed from the information security community is to only have the data you need for the time when you need it. You may be tempted to keep your data forever just in case you need it someday. However, the technology, data and political landscape is constantly shifting, and you don't know if it will be possible to use the information to harm someone or some group in the future. If it is important to keep the data forever, you must plan for resources to continually update and support safe data management practices for the data.

disposing data

The key to disposing your information is knowing where your data is in order to delete all of it (see above on 'what data do I have and where is my data').

Deleting digital information is more than just clicking "delete", but it is not hard to do. As a general rule, everything that is on a hard drive needs to be overwritten several times (there are free tools to help you do this) -EXAMPLE? . But if your needs are more specific, do not hesitate to reach out to an expert to help you navigate this.

archiving data

Archiving is a general term for the range of practices and decisions that support the long-term preservation, use, and accessibility of content with enduring value. It is not a one-time action, but is instead a process and an investment that connects directly to your projects' goals.

should you archive your data?

You may want to archive your information if it has enduring cultural, historical, or evidentiary value. Preserving information has the potential to support the protection of rights, to seek redress, and to support reconciliation or recovery in damaged societies.

planning for your data to be archived

Try to identify as early as possible that you want archive, what you want to archive and where you want to archive. When you are at the closing stage, it may be overwhelming to solve all arising problems, while having limited time before you need to move on.

Ideally you will identify the metadata that you need about the information you are collecting so that you don't need to collect too much and you don't collect too little.

Try to store data to be archived in popular (=interoperable) formats which are likely to be used in future (for example, they are used by popular software). Over the course of a few years formats may go out of use and software systems become unable to use the data.

Keep in mind that videos can be very large, so you need to make sure that you have the infrastructure or support in place to accommodate this data (see potential archive partners below).

where do you archive?

If you archive data yourself, you will need to consider the cost, time and skills that will be required to maintain the archival system. This is not for the faint of data-heart! See (and learn intimately) the [SECTION ON STORAGE] and include data storage capacity into your long-term organisational strategy.

Alternatively, it might be useful to partner directly with an archive that will help you in the public interest, such as:

> - The Open Society Archives http://www.osaarchivum.org/
> - Duke University Human Rights Archive
> http://library.duke.edu/rubenstein/human-rights/
> - Human Rights Web Archive at Columbia University
> http://hrwa.cul.columbia.edu/
> - University of Texas Libraries' Human Rights Documentation Initiative
> http://www.lib.utexas.edu/hrdi
> - Archivists without Borders http://www.arxivers.org/en/asf_internacional.php

further resources

- Archivists' Guide to Archiving Video (WITNESS) http://archiveguide.witness.org/ and a video on this topic http://blog.witness.org/2014/10/video-series-archiving-and-preservation-activists/
- New Tactics online discussion on archiving for human rights advocacy, justice and memory https://www.newtactics.org/conversation/archiving-human-rights-advocacy-justice-and-memory

additional resources

getting data resources

understanding data resources

sharing data resources

existing data resources

organizations to reach out to for urgent support

project design resources

data management resources

getting data resources

- Conducting safe, effective and ethical Interviews with survivors of Sexual and Gender-Based
Violence (WITNESS) http://www.scribd.com/doc/159998474/Conducting-Safe-Effective-and-Ethical-Interviews-with-Survivors-of-Sexual-and-Gender-Based-Violence
- Draft Principles for Resilliance and Big Data (Bellagio/PopTech Fellows) - The following is a draft “Code of Conduct” that seeks to provide guidance on best practices for resilience building projects that leverage Big Data and Advanced Computing. http://irevolution.net/2013/09/23/principles-for-big-data-and-resilience/
- Code of Ethics for the Proper Use of Social Data (Big Boulder Initiative) - An effort to begin to define a set of ethical values and practices for the treatment of social data, and to educate the industry about ethical social data collection, processing and utilization
practices. http://blog.bigboulderinitiative.org/2014/06/05/a-code-of-ethics-for-social-data-we-need-your-help/
- Big data: Philanthropy or privacy invasion? by Ayee Macaraig
http://www.rappler.com/world/regions/asia-pacific/42116-big-data-philanthropy-privacy-invasion
- Open Data & Privacy Discussion Notes (Open Data Research Network)
http://www.opendataresearch.org/content/2013/501/open-data-privacy-discussion-notes
- Open Data Impacts Research Blog (Tim Davies)
http://www.opendataimpacts.net/
- Big Data and Privacy (NYU School of Law) reducing predictive harm
http://lsr.nellco.org/cgi/viewcontent.cgi?article=1434&context=nyu_plltwp
- Guidelines on use of video for documentation, including human rights abuses (Video 4 Change) https://www.v4c.org/category/safety-and-security
- Ethical Decision-Making and Internet Research
http://aoir.org/reports/ethics2.pdf
- Internews - SpeakSafe: Media Workers’ Tool for Safer Online and Mobile Practices - https://internews.org/research-publications/speaksafe-media-workers-toolkit-safer-online-and-mobile-practices

understanding data resources

- A gentle introduction to cleaning data (School of Data) http://schoolofdata.org/handbook/courses/data-cleaning/
- A gentle introduction to exploring and understanding your data (School of Data) http://schoolofdata.org/handbook/courses/gentle-introduction-exploring-and-understanding-data/
- Data QualYtI: Do you trust your data? (an article Hjusein Tjurkmen, Mariyana Hristova, Musala Soft) http://istabg.org/data-qualyti-do-you-trust-your-data/
- Open Refine offers powerful ways to combine, compare and reconcile your data. http://openrefine.org/
- The Storyful Blog and Case Studies provide a constant stream of up-to-date information on the ever-challenging video verification process http://blog.storyful.com/
- Verification Handbook is authored by leading journalists from the BBC, Storyful, ABC, Digital First Media and other verification experts. Created for journalists and human rights defenders as it provides the tools, techniques and step-by-step guidelines for how to deal with user-generated content (UGC) during emergencies. http://verificationhandbook.com/
- Truth in the Age of Social Media (The Nieman Lab report) http://nieman.harvard.edu/reports/issue/100072/Summer-2012.aspx
- Citizen Evidence Lab - tools for speedy checks on YouTube videos as well as for more advanced assessment http://citizenevidence.org/
- Authenticating Open Source Video (WITNESS) http://www.mediafire.com/view/cd7b76ydpe22khn/AuthenticatingOpenSourceVideo_2013.pdf
- Data Integrity User Guide (FrontlineSMS) - The guide is intended to help users to understand, analyze, and address the vulnerabilities, risks and threats that can affect the integrity of the information communicated through the FrontlineSMS platform. http://www.frontlinesms.com/wp-content/uploads/2011/08/frontlinesms_userguide.pdf

sharing data resources

- A Qualitative Risk Assessment Framework for Sharing Computer Network Data (Scott E. Coull RedJack Erin Kenneally) http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2032315.
- Handbook on Statistical Disclosure Control (CNEX project) http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
- Heuristics for de-identifying health data (Khaled el emam Children's Hospital of Eastern Ontario Research Institute) http://ruor.uottawa.ca/fr/bitstream/handle/10393/12987/El_Emam_Khaled_2008_Heuristics_for_de-identifying_health_data.pdf?sequence=1
- Choosing an Open Source Licence for code - http://choosealicense.com
- Content Licencing https://creativecommons.org

existing data policies and guidelines

- InterAction Protection Working Group: Data Collection in Humanitarian Response - A Guide for Incorporating Protection http://pqdl.care.org/Practice/Data%20Collection%20in%20Humanitarian%20Response,%20A%20Guide%20for%20Incorporating%20Protection.pdf
- Post about forthcoming OCHA report on Humanitarianism in the age of Cyber Warfare - Towards the Principled and Secure Use of Information in Humanitarian Emergencies http://wget2014.wordpress.com/2014/04/22/15-humanitarianism-in-the-cyberwarfare-age-the-principled-and-secure-use-of-information-in-humanitarian-response/ (April 2014)
- Guidance for Establishing Affected Persons Information Center (Viktoria Lovrics and Andrej Verity (UN Office of the Coordination of Humanitarian Affairs (OCHA)) http://blog.veritythink.com/post/98874290694/guidance-for-establishing-an-affected-persons
- International Committee of the Red Cross (ICRC) - Professional standards for protection work carried out by humanitarian and human rights actors in armed conflict and other situations of violence https://www.icrc.org/eng/resources/documents/publication/p0999.htm
- 2013 OECD Privacy Guidelines http://www.oecd.org/sti/ieconomy/privacy.htm
- Guidelines for Secure Use of Social Media by Federal Departments and Agencies (CIO Council) - This guidelines document is written for the US government and yet its description of the benefits and risks of using social media and networking sites are in part appropriate to the needs and risks of NGOs https://cio.gov/wp-content/uploads/downloads/2012/09/Guidelines_for_Secure_Use_Social_Media_v01-0.pdf
- Applying Ethical Principles to Information and Communication Technology Research: A Companion to the Department of Homeland Security Menlo Report http://www.cyber.st.dhs.gov/wp-content/uploads/2012/01/MenloPrinciplesCOMPANION-20120103-r731.pdf
- NYU School of Law - Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms http://lsr.nellco.org/cgi/viewcontent.cgi?article=1434&context=nyu_plltwp

- Children's Online Privacy and Protection Act (COPPA) http://www.coppa.org/
- ICRC Professional Standards for Protection Work **-** These standards are addressed to all humanitarian and human rights actors engaged in protection work in favour of communities and persons at risk in armed conflict and other situations of violence. Chapter 6 page 37 is on 'Managing sensitive protection information' https://www.icrc.org/eng/assets/files/other/icrc-002-0999.pdf
- IOM Data Protection Manual - The manual is comprised of three parts: the first part outlines the IOM data protection principles as informed by relevant international standards; the second part includes comprehensive guidelines on each principle, consideration boxes and practical examples; and the third part provides generic templates and checklists to ensure that data protection is taken into account when collecting and processing personal data
- http://publications.iom.int/bookstore/index.php?main_page=product_info&cPath=47&products_id=759
- Ethical Framework for Researchers Using and Collecting Data on the M-Lab Platform for Mobile Connectivity Measurements (Oxford Internet Institute)
- http://www.oii.ox.ac.uk/research/projects/?id=107
- Donor code of conduct http://www.ssireview.org/blog/entry/a_new_donor_code_of_conduct
- UN data collection: http://www.unglobalpulse.org/privacy-and-data-protectionwhy is it important
- Professional standards for protection work carried out by humanitarian and human rights actors in armed conflict and other situations of violence: http://www.icrc.org/eng/resources/documents/publication/p0999.htm
- UNFPA guidelines on data issues in Humanitarian Crisis situations https://www.unfpa.org/public/home/publications/pid/6253
- Protection International – Security Manual Updating - http://protectioninternational.org/publication/new-protection-manual-for-human-rights-defenders-3rd-edition/

resources

organizations to reach out to for urgent support

HIVOS Digital Defenders Emergency Response and Grants - http://digitaldefenders.org

If you are a human rights defender, journalists, blogger, activist, NGO or media organization and you need immediate help to mitigate a digital emergency please click here [https://digitaldefenders.org/]. If you think something is wrong with your computer, phone or accounts (email, social media, website or other) please get in contact with us. See also: Digital First Aid Kit http://digitaldefenders.org/digitalfirstaid

freeDimensional - http://freedimensional.org/services/distress-services/

Distress Services are intended for activists and culture workers in situations of distress as a result of their professional work.

Frontline Defenders - http://www.frontlinedefenders.org/emergency

Front Line seeks to provide 24 hour support to human rights defenders at immediate risk. If there is a crisis you can contact Front Line at any hour on the emergency hotline.

Urgent Action Fund for Women's Human Rights - http://urgentactionfund.org/apply-for-a-grant/apply-for-an-evacuation-grant/

Urgent Action Fund supports women activists who are being threatened because of their work defending human rights. The Evacuation Grant is a specific type of Rapid Response Grant, designed for those in urgent need of relocation funding because of threats, persecution and/or an extreme security situation.

project design resources

- Privacy Impact Assessments Handbook (Information Commissioner's Office (UK)) http://www.ico.org.uk/upload/documents/pia_handbook_html_v2/files/PIAhandbookV2.pdf
- Security in a Box https://securityinabox.org/ and Me and my shadow https://myshadow.org/ (Tactical Technology Collective)
- Dialing Down Risks: Mobile Privacy and Information Security in Global Development Projects (New America Foundation) http://newamerica.net/publications/policy/dialing_down_risks_mobile_privacy_and_information_security_in_global_development
- On-going study into digital security for aid agencies (European Inter-agency Security Forum) http://www.eisf.eu/resources/home.asp
- Online and Physical Security Training (Save the Children) http://www.disasterready.org/

data management resources

- Deidientification Maturity Model (Privacy Analytics) http://waelhassan.com/wp-content/uploads/2013/06/DMM-Khaled-El-Emam-Wael-Hassan.pdf
- Guide to Protecting the Confidentiality of Personally Identifiable Information (National Institute of Standards and Technology) http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf
- The Frontline SMS Users' Guide to Data Integrity http://www.frontlinesms.com/wp-content/uploads/2011/08/frontlinesms_userguide.pdf
- For a listing of secure tools, see https://www.prismbreak.org
- On choosing a hosting provider, see https://learn.equalit.ie/wiki/Responsible_Data_Forum_on_Hosting
- On setting up a secure hosting provider, see https://learn.equalit.ie/wiki/Secure_hosting_guide
- Tactical Tech's Security in a Box chapter on protecting physical assets https://securityinabox.org/chapter-2
- OECD Privacy Principles http://oecdprivacy.org/
- Anonymisation: managing data protection risk code of practice (Information Commissioner's Office UK) http://www.ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation